UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

Mejorando la evaluación de juegos serios aplicando analíticas de aprendizaje y técnicas de minería de datos

Improving serious games evaluation by applying learning analytics and data mining techniques

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Cristina Alonso Fernández

DIRECTORES

Baltasar Fernández Manjón
Manuel Freire Morán
Iván Martínez Ortiz

Madrid, 2021

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

MEJORANDO LA EVALUACIÓN DE JUEGOS SERIOS APLICANDO ANALÍTICAS DE APRENDIZAJE Y TÉCNICAS DE MINERÍA DE DATOS
IMPROVING SERIOUS GAMES EVALUATION BY APPLYING LEARNING ANALYTICS AND DATA MINING TECHNIQUES

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

CRISTINA ALONSO FERNÁNDEZ

DIRECTOR

BALTASAR FERNÁNDEZ MANJÓN
MANUEL FREIRE MORÁN
IVÁN MARTÍNEZ ORTIZ

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

MEJORANDO LA EVALUACIÓN DE JUEGOS SERIOS APLICANDO ANALÍTICAS DE APRENDIZAJE Y TÉCNICAS DE MINERÍA DE DATOS

IMPROVING SERIOUS GAMES EVALUATION BY APPLYING LEARNING ANALYTICS AND DATA MINING TECHNIQUES

CRISTINA ALONSO FERNÁNDEZ

DIRECTORES

BALTASAR FERNÁNDEZ MANJÓN
MANUEL FREIRE MORÁN
IVÁN MARTÍNEZ ORTIZ

Madrid, 2021

"I am never so happy as when I am really engaged in good earnest,
and it makes me must wonderfully cheerful and merry at other times,
which is curious and very satisfactory"

Ada Lovelace

Acknowledgments

From the outside, it could seem that "the thesis" is mainly this document but, actually, it is all the work carried out during, at least, the last 4 years. During this time, I have been lucky to have the support and company of a lot of people that have made this journey, that is sometimes confusing, much more bearable.

Thanks to my classmates *doblegradistas* and of the master, as well as to the teachers that I had during those years, that helped me to lay the groundwork for this adventure. Thanks to Rafa for his help and co-direction of my Final Master Thesis.

Thanks to the Department of Software Engineering and Artificial Intelligence, in which I have carried out this work. Special mention to Lourdes, for her help and joy. Thanks to the members of the Doctoral Program, specially Narciso, Román and Daniel for their help. Thanks to the staff of the Computer Science Faculty, specially to Sánchez for his red-hot coffees that woke us up even in the most tiring afternoons.

Thanks to my workmates in the research stay in Florida State University: mainly Professor Valerie Shute, who always helped me. And her research group, with whom I had the luck to work a few months: Ahmad, Lukas, Ginny, Xiaotong, Renata, Chi-Puh, Curt, Chen, Russell.

Thanks to my mates of the room 16 and surroundings for their company while working and, specially, while not working: Toni, Iván, Victorma, Cristian, Marta, Pablo, Miguel, Alicia, Jesús, Luisma, Joaquín, Dani, Rubén. Although we have repeated this many times, it is still true: without you, I would have finished this thesis much earlier, but it would have been a lot less fun.

Thanks to the other members of the e-UCM research group who helped me during this work: Ángel, Ana, Alma. Thanks to Julio for his work during his stay.

Thanks to my supervisors for their advice and constant help during these years. To Iván, *Lord of the Machines*, and Manu, *Master of English*, for their different perspectives that have enriched this work. And, of course, to Balta, for his guidance and help all these years, since I had the crazy idea of asking him to supervise my Final Degree Project.

Thanks to my friends and family, who, although they did not always understand what I was doing, always supported me in this journey.

And mostly, thanks to my mother and father, for cheering me up in the bad moments, for accompanying me in the good ones, and for everything else.

Agradecimientos

Desde fuera, podría parecer que "la tesis" se compone principalmente de este documento, pero, en realidad, es el cúmulo del trabajo realizado durante, al menos, los últimos 4 años. En este tiempo, he contado con el apoyo y la compañía de muchas personas que han hecho este camino, en ocasiones confuso, mucho más transitable.

Gracias a mis compis *doblegradistas* y del máster, así como a los profesores y profesoras que tuve durante esos años, que me ayudaron a sentar las bases en las etapas previas a esta aventura. Gracias a Rafa por su ayuda y codirección del Trabajo de Fin de Máster.

Gracias al equipo del Departamento de Ingeniería del Software e Inteligencia Artificial, en el que he realizado este trabajo. Mención especial a Lourdes, por su ayuda y alegría. Gracias al equipo del Programa de Doctorado, en especial a Narciso, Román y Daniel por su ayuda. Gracias al personal de la Facultad de Informática, en especial a Sánchez por sus cafés *al rojo vivo* que nos espabilaban hasta en las tardes más cansadas.

Gracias a las personas con las que pasé la estancia en la Universidad Estatal de Florida: principalmente a la Profesora Valerie Shute, que me ayudó en todo momento. Y a todo su grupo de investigación con los que tuve la suerte de trabajar unos meses: Ahmad, Lukas, Ginny, Xiaotong, Renata, Chi-Puh, Curt, Chen, Russell.

Gracias a mis compis del aula 16 y alrededores por su compañía en los momentos de trabajo y, sobre todo, en los de no trabajo: Toni, Iván, Victorma, Cristian, Marta, Pablo, Miguel, Alicia, Jesús, Luisma, Joaquín, Dani, Rubén. Aunque lo hayamos repetido muchas veces, sigue siendo verdad: sin vosotros hubiera terminado esta tesis mucho antes, pero hubiera sido mucho menos divertido.

Gracias también al resto del grupo e-UCM que me ha ayudado en este trabajo: Ángel, Ana, Alma. Gracias a Julio por su colaboración en los meses de su estancia.

Gracias a mis directores por sus consejos y ayuda constante en estos años. A Iván, *señor de las máquinas*, y a Manu, *master of English*, por sus puntos de vista tan complementarios para enriquecer el trabajo. Y, por supuesto, a Balta, por su guía y ayuda todos estos años, desde que tuve la loca idea de pedirle que me dirigiera el Trabajo de Fin de Grado.

Gracias a mis amigas y amigos, y a mi familia, que, aunque no siempre entendieran a que dedicaba el tiempo, han estado siempre apoyándome en este camino.

Y sobre todo gracias a mi madre y a mi padre, por animarme en los momentos malos, por acompañarme en los momentos buenos, y por todo lo demás.

About this document

The thesis presented in this document was carried out as a compendium of publications. The publications included in the thesis are listed below, and their full text is included in Chapter 6.

Journal publications:

- Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2019): **Applications of data science to game learning analytics data: a systematic literature review.** Computers & Education, Volume 141, November 2019, 103612. DOI: 10.1016/j.compedu.2019.103612.
    - Impact metrics: JCR 2019, Impact Factor: 5.296, Q1 in Computer Science, Interdisciplinary Applications.
    - This paper presents the systematic literature review carried out about the applications of data mining techniques to game learning analytics data from serious games.
    - Details and results of the paper are described in Section 2.3 of this document, as part of the presentation of the state of the art, and in Section 4.1, as part of the results of the thesis.
- Cristina Alonso-Fernández, Iván Martínez-Ortiz, Rafael Caballero, Manuel Freire, Baltasar Fernández-Manjón (2020): **Predicting students' knowledge after playing a serious game based on learning analytics data: A case study.** Journal of Computer Assisted Learning, vol. 36, no. 3, pp. 350-358, June 2020. DOI: 10.1111/jcal.12405.
    - Impact metrics: JCR 2019, Impact Factor: 2.126, Q2 in Education & Educational Research.
    - This paper presents the first case study carried out to test our assessment approach using learning analytics data and prediction models.
    - Details and results of the paper are described in Section 4.2 of this document.
- Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2020): **Evidence-based evaluation of a serious game to increase bullying awareness.** Interactive Learning Environments, 2020. DOI: 10.1080/10494820.2020.1799031.
    - Impact metrics: JCR 2019, Impact Factor: 1.938, Q2 in Education & Educational Research.

- This paper presents the second case study carried out to further explore our assessment approach with a different serious game.
- Details and results of the paper are described in Section 4.3 of this document.

- Cristina Alonso-Fernández, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2021): **Improving evidence-based assessment of players using serious games.** Telematics and Informatics. (in press) DOI: 10.1016/j.tele.2021.101583.
  - Impact metrics: JCR 2019, Impact Factor: 4.139, Q1 in Information Science & Library Science.
  - This paper presents the final evidence-based assessment process of serious games players based on game learning analytics data and prediction models.
  - Details and results of the paper are described in Section 4.4 of this document.
- Cristina Alonso-Fernández, Ana Rus Cano, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2019): **Lessons learned applying learning analytics to assess serious games.** Computers in Human Behavior, Volume 99, October 2019, Pages 301-309. DOI: 10.1016/j.chb.2019.05.036.
  - Impact metrics: JCR 2019, Impact Factor: 5.003, Q1 in Psychology, Experimental.
  - This paper presents the research carried out in this and two other thesis exploring different applications of learning analytics data to assess serious games.
  - Details and results of the paper are described in Section 4.5 of this document.

Conference publications:

- Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2017): **Systematizing game learning analytics for serious games.** IEEE Global Engineering Education Conference (EDUCON), 25-28 April 2017, Athens, Greece.
  - This paper presents the first steps carried out in systematization of the application of game learning analytics in serious games.
  - Details and results of the paper are described in Section 4.4 of this document.

    - This paper received a **Best Paper Award** of the Conference, in the "Area 3: Innovative Materials, Teaching and Learning Experiences in Engineering Education".
- Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2021): **Data science meets standardized game learning analytics.** IEEE Global Engineering Education Conference (EDUCON), 21-23 April 2021, Vienna, Austria.
    - This paper presents the tool T-MON, an analysis and visualization tool to conduct exploratory analysis on the game interaction data collected.
    - Details and results of the paper are described in Section 4.4 of this document.
- Cristina Alonso-Fernández, Dan C. Rotaru, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2017): **Full Lifecycle Architecture for Serious Games: Integrating Game Learning Analytics and a Game Authoring Tool.** Joint Conference on Serious Games (JCSG), 23-24 November 2017, Polytechnic University of Valencia, Spain.
    - This paper presents the work to integrate game learning analytics data as part of a game authoring tool.
    - Details and results of the paper are described in Section 4.5 of this document.
- Cristina Alonso-Fernández, Ivan Perez-Colado, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2018): **Improving serious games analyzing learning analytics data: lessons learned.** Games and Learning Alliance conference (GALA Conf), December 5-7, 2018, Palermo, Italy.
    - This paper presents the initial lessons learned in some of the work carried out with learning analytics data in serious games.
    - Details and results of the paper are described in Section 4.5 of this document.
- Cristina Alonso-Fernández, Ana Rus Cano, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2018): **Applications of learning analytics to assess serious games**. 2nd Annual Learning & Student Analytics Conference (LSAC), October 22-23, 2018, Amsterdam, The Netherlands.
    - This paper explored the opportunities for assessment with learning analytics in serious games.
    - Details and results of the paper are described in Section 4.5 of this document.

Abstract

Title: Improving serious games evaluation by applying learning analytics and data mining techniques.

Serious games are highly motivational resources effective to teach, raise awareness, or change the perceptions of players. To foster their application in education, teachers and institutions require clear and formal evidences to assess students' learning while they are playing the games. However, traditional assessment techniques rely on external questionnaires, typically carried out before and after playing, that fail to measure players' learning while it is happening. The multiple interactions carried out by players in the games can provide more precise information about how players play, and even be used to assess them. In this regard, game learning analytics techniques propose the collection and analysis of such interactions for multiple purposes, including assessment. The potentially large game learning analytics data collected can be further analyzed with data mining techniques to discover unexpected patterns and to provide measures to evaluate the effect of games on their players and assess their learning.

In this thesis, we propose a new approach to assess serious games players based on evidences collected from their gameplays. The interaction data collected is analyzed to derive game learning analytics variables that fill data mining models to predict players' learning. These prediction models are created during the serious games' validation phase and tested by comparing their predictions to the learning measured by the differences between the pre-post questionnaires. With our approach, once the games and the prediction models are validated, players' assessment in large-scale deployment is simplified as players can be assessed solely based on their interaction data, as questionnaires are no longer required. The approach uses a standard data collection format for interactions with serious games, xAPI-SG, to systematize the interaction data collected, as well as the creation of the relevant game learning analytics variables. Finally, to support the essential step of creating those variables, we provide an exploratory tool, T-MON, to analyze and visualize the collected interaction data.

The approach is based on the lessons learned in two case studies with different serious games in which we conducted all the steps to assess students based on their interactions. Simplifying and improving players' assessment, educators and institutions will have clearer evidences to include serious games in their classes, without the

additional costs for players to complete the questionnaires, and for educators to then analyze their results. Moreover, with this approach, we aim to contribute to the systematization of players' assessment, one of the gaps identified in the systematic literature review about data science applications to game learning analytics data.

The work carried out in this thesis builds on the work I performed as a member of the e-UCM group, including work conducted as part of the H2020 European projects RAGE and BEACONING. During those projects, I worked in improving the application and deployment of serious games and game learning analytics in large scale deployment scenarios, dealing with both the technological part as well as the analysis of the interaction data collected. Moreover, in the final part of the thesis, during a research stay at Florida State University at Professor Valerie Shute's research group in *stealth assessment*, I worked in an international environment using the hosting group's game learning analytics and player assessment techniques, which allowed me to do an initial validation and contrast the work presented in this proposal.

Keywords: *serious games, learning analytics, game-based learning, stealth assessment, data mining, standardization, e-learning*

Resumen

Título: Mejorando la evaluación de juegos serios aplicando analíticas de aprendizaje y técnicas de minería de datos.

Los juegos serios son recursos altamente motivadores y efectivos para enseñar, concienciar, o cambiar las percepciones de sus jugadores. Para fomentar su aplicación en educación, los profesores y las instituciones necesitan pruebas claras y automáticas con las que evaluar el aprendizaje de sus estudiantes mientras utilizan los juegos. Tradicionalmente, la evaluación con juegos serios se basa en cuestionarios externos, realizados normalmente antes y después de jugar, que no miden el aprendizaje de los jugadores durante el proceso en sí. Las múltiples interacciones que realizan los jugadores al jugar pueden proporcionar una información más precisa sobre cómo juegan los jugadores e, incluso, utilizarse para evaluar su aprendizaje. En este sentido, las analíticas de aprendizaje para juegos proponen técnicas para la recogida y el análisis de dichas interacciones con múltiples fines, incluida la evaluación de los jugadores. Los datos (potencialmente numerosos) de las analíticas de aprendizaje para juegos pueden analizarse en mayor detalle con técnicas de minería de datos que permiten descubrir patrones ocultos a simple vista y proporcionar mejores medidas para estudiar el efecto de los juegos en los estudiantes y evaluar su aprendizaje.

En esta tesis, proponemos un nuevo método para evaluar a los jugadores de juegos serios basándonos en las evidencias recogidas mientras juegan. Los datos de interacción recogidos se analizan para extraer variables de analíticas de aprendizaje utilizadas por modelos predictivos de minería de datos para cuantificar el aprendizaje de los jugadores. Estos modelos predictivos se crean durante la fase de validación de los juegos serios, y se validan comparando sus predicciones con el aprendizaje medido por las diferencias entre los cuestionarios anterior y posterior al juego. Con nuestra propuesta, una vez validados los juegos y los modelos predictivos, la evaluación de los jugadores se simplifica durante el despliegue a gran escala, permitiendo que los jugadores pueden ser evaluados automáticamente con sus datos de interacción, sin necesidad de cuestionarios. La propuesta utiliza un formato de recogida de datos estándar para las interacciones con juegos serios, xAPI-SG, que permite sistematizar tanto los datos de interacción recogidos, como la creación de variables de analíticas de aprendizaje. Por último, para ayudar en la etapa esencial de extracción de variables, proporcionamos una herramienta exploratoria, T-MON, para analizar y visualizar los datos de interacción recogidos.

La propuesta se basa en las lecciones aprendidas en dos casos de estudio con diferentes juegos serios en los que realizamos todos los pasos para evaluar a los estudiantes basándonos en sus interacciones. Simplificando y mejorando la evaluación de los jugadores, los educadores y las instituciones tendrán evidencias más claras para incluir los juegos serios en sus clases, sin los costes adicionales para los jugadores de completar dichos cuestionarios, y para los educadores de analizarlos posteriormente. Además, con esta propuesta buscamos avanzar en la sistematización de la evaluación de los jugadores, uno de los vacíos identificados en la revisión sistemática de la literatura sobre las aplicaciones de ciencia de datos a analíticas de aprendizaje para juegos.

El trabajo realizado en esta tesis se basa en el trabajo en el que he participado como integrante del grupo e-UCM, incluyendo la investigación que formó parte de los proyectos europeos H2020 RAGE y BEACONING. Durante estos proyectos, trabajé en mejorar la aplicación y el despliegue de juegos serios y de analíticas de aprendizaje para juegos en entornos de desarrollo de gran escala, abordando tanto la parte de tecnologías aplicadas como la parte de análisis de los datos de interacción recogidos. Además, durante la parte final de la tesis, realicé una estancia de investigación en la Universidad Estatal de Florida, con el grupo de investigación en *stealth assessment* de la Profesora Valerie Shute, en la que trabajé en un entorno internacional con técnicas de analíticas de aprendizaje y evaluación de estudiantes, que me permitieron realizar una validación inicial y contrastar el trabajo que presento en esta propuesta.

Palabras clave: *juegos serios, analíticas de aprendizaje, aprendizaje basado en juegos, evaluación, minería de datos, estandarización, e-learning*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADL	Advanced Distributed Learning
AI	Artificial Intelligence
API	Application Programming Interface
CV	Cross Validation
DM	Data Mining
DSRM	Design Science Research Methodology
EDM	Educational Data Mining
EU	European Union
e-UCM	UCM research group on e-learning technologies
FN	False Negative
FP	False Positive
GA	Game Analytics
GDPR	General Data Protection Regulation
GLA	Game Learning Analytics
IEEE	Institute of Electrical and Electronics Engineers
JSON	JavaScript Object Notation
k-NN	K-Nearest Neighbors
LA	Learning Analytics
LAM	Learning Analytics Model
LMS	Learning Management System
MAE	Mean Absolute Error
MOOC	Massive Open Online Course
MR	Misclassification Rate
RQ	Research Question

SD	Standard Deviation
SG	Serious Game
SVM	Support Vector Machines
SVR	Support Vector machines for Regression
TN	True Negative
TP	True Positive
T-MON	Monitor of xAPI-SG traces
UCM	Universidad Complutense de Madrid
xAI	Explainable Artificial Intelligence
xAPI	Experience API
xAPI-SG	Experience API for Serious Games

Chapter 1. Introduction

This chapter presents the motivation for the work carried out in this thesis, namely the improvement of the assessment method of students playing serious games, taking advantage of game learning analytics data and its analysis with data mining techniques. The chapter further summarizes the structure of the following chapters of this document.

1.1. Motivation

Serious games are videogames that aim to cause an effect on players beyond simple entertainment, for example, increase players' knowledge or awareness about social issues. We consider that serious games offer new opportunities not only as tools of learning but, additionally, as tools to assess that learning due to their multiple advantages, from their interactive and engaging nature to the possibility to test complex scenarios in a safely manner. Nevertheless, assessment of serious games effect on players is still conducted mostly through external formal questionnaires, which merely assess players before and after the gameplay, missing the opportunity to perform the assessment while players are actually learning, i.e., while they are playing the game.

The innovative field of Game Learning Analytics proposes the collection, analysis and report of information extracted from the data obtained from players' interactions with serious games. The application of Game Learning Analytics brings new opportunities to improve different aspects of the serious games lifecycle: further explore players' behaviors in the gameplays, understand their progress and learning processes, improve the game and learning design, and adapt the learning experience to players' characteristics and needs. The rich and potentially large amount of data gathered from the collection of game learning analytics can be further analyzed with complex analysis techniques to discover unexpected patterns. Data mining techniques like prediction models provide further opportunities to analyze the collected data, and together with the insight provided by game learning analytics, can allow to perform players' assessment using serious games in a precise and automatic way without relying on external questionnaires. Accurate prediction models could be trained to automatically assess players solely based on their game interaction data.

Such data-based assessment of players can greatly simplify the process to obtain evidences on how much serious games are impacting their players. The simplification of the students' assessment can provide the required tools to verify that students are learning, therefore, increasing teachers and institutions trust in serious games as tools for causing an actual effect on students/players. This, in turn, can contribute to expand the application of serious games in real life scenarios – beyond their current limited application as a complementary activity with no actual impact on students' evaluations. Therefore, the simplification of the current assessment techniques of players using serious games can be one of the factors that contribute to foster the application of serious games.

In our previous work related to serious games, we had analyzed the potential of game learning analytics data gathered from players' interactions. Our research group e-UCM was part of two H2020 European projects, RAGE and BEACONING, in which, among other tasks, we managed the game learning analytics collection, analysis and reporting. During those projects, we worked in improving the application and the large-scale deployment of serious games, collecting and analyzing large amounts of game learning analytics data. For that work, the application of an interaction data standard format (xAPI-SG) was essential to provide a clear definition of the data collected from the different serious games and to systematize the analysis and visualizations carried out. This was particularly important working in an international project, with multiple partners that had to collaborate in the development of the tools. The work carried out in those projects laid the foundation for this work, providing a valuable experience in large-scale deployment of serious games in real scenarios.

Additionally, in the final stages of this work, I had the opportunity to carry out a research stay in Florida State University with Professor Valerie Shute and her group, a leading research group in the field of *stealth assessment*. During that stay, besides learning the process and methodology of the work that they carry out, we worked together in two studies applying different techniques to the game learning analytics data collected from a serious game that they were studying. This experience provided me with a great insight on their techniques and methodology, as well as with a different perspective on the assessment of players with serious games. In the studies conducted, we were able to apply some of the steps and processes of our assessment approach, presented in this thesis, in a different context (different serious game and interaction data format), validating some of the steps of our approach and obtaining additional results on the application of game learning analytics in serious games.

1.2. Document structure

The rest of this document is structured as follows:

- Chapter 2 reviews the state of the art about the three main topics that are central to the thesis work: Serious Games, including the current methods of assessment of their players; Game Learning Analytics, including the standardization of the data collected (and the xAPI-SG Profile); and data mining techniques, such as prediction models. The chapter concludes analyzing the combination of the three previous topics, that is, the application of data mining techniques to game learning analytics data from serious games, including the results obtained in the systematic literature review conducted about this application.
- Chapter 3 states the research goals of the thesis and presents the research methodology carried out and the steps followed in the thesis.
- Chapter 4 presents the process, results and conclusions obtained in the systematic review of the literature, in the two case studies carried out with different serious games to conduct evidence-based assessment of their players, and the final evidence-based assessment process obtained, detailing all the steps needed, the data standard use, and presenting tools to support the process. The chapter concludes with a discussion of the work carried out and its limitations.
- Chapter 5 summarizes the conclusions obtained from our work, the main contributions of the thesis and some of the possible lines of future research.
- Finally, Chapter 6 contains the details and full text of the journal and conference publications that constitute the thesis.

Chapter 2. State of the art

This chapter summarizes the state of the art in relation to: serious games, including the techniques to assess their players; game learning analytics, and data standardization, including the data standard for interactions with serious games xAPI-SG; and data mining techniques, including prediction models. The chapter finally presents the combination of the three previous fields that compose the core of the thesis: the applications of data mining techniques to game learning analytics data collected from serious games, including the systematic literature review carried out about this topic, that constitutes one of the contributions of the thesis.

2.1. Serious Games

The application of innovative and more interactive tools in educational contexts has greatly spread in recent years. For instance, the field of e-learning proposes a learning experience through electronic tools including more means of interaction, although sometimes the experience is restricted to a digitalized version of traditional learning. Gamification techniques (i.e. the use of game techniques in non-gaming areas) are also been applied in education to benefit from their advantages compared to more traditional approaches. This interest has also reached videogames, and their application with serious purposes has greatly increased in recent years.

Serious Games (SGs) have been defined as games that "do not have entertainment, enjoyment or fun as their primary purpose" (Michael & Chen, 2005) and as digital games "created with the intention to entertain and to achieve at least one additional goal (e.g., learning or health)" (Dörner, Göbel, Effelsberg, & Wiemeyer, 2016). Although serious games also should be entertaining, their main purpose could be to teach some knowledge, create awareness of some issue, change players' attitudes, etc.

Videogames provide an engaging, highly interactive environment with many possibilities for causing an effect on players. Among their several advantages, they provide (Dörner et al., 2016):

- **Motivation**: games increase students' motivation, allowing them to interact with the learning tool and overcoming the proposed challenges.

- **Engagement**: games reach players on an emotional level, in an immersive experience, breaking the usual barrier of 10 minutes of attention. This way, serious games allow to link education and entertainment.
- **Feedback and adaptation**: within games, players can practice different strategies or choices, obtaining feedback of the consequences of their actions. Games also provide adaptability, as they can change according to players' choices or profiles.
- **Progress and completion**: games provide a progressive increase in the difficulty so players can train the skills or knowledge, while they are progressing in the game. They also provide a means of completion to the full game or different levels or chapters that provide a feeling of progress within the gameplay.
- **Free and safe exploration**: serious games allow players to test complex or risky scenarios in a safely manner, as players can explore the game areas and path in a safe and free exploration, training complex procedures.
- **Active learning**: the interactive nature of games helps in allowing students to have an active role for learning; compare to the traditional passive role of learning in traditional lectures.

Examples of Serious Games can be found in multiple domains, such as: medicine (Evans et al., 2015; Standford Medicine, 2013), mathematics (Center for Game Science at the University of Washington, 2016), physics (V. J. Shute, Ventura, & Kim, 2013), literature (Iglesias, Fernandez-Vara, & Fernandez-Manjon, 2013), history (GTLHistory, 2020), computer science (Adamo-Villani, Haley-Hermiz, & Cutler, 2013), or military (United States Army, 2002). Besides teaching knowledge, other serious games focus on raising awareness about social problems, such as humanitarian crisis (interFUEL, 2006) or drug addiction (Asociación Servicio Interdisciplinar de Atención a las Drogodependencias (SIAD), 2014). Some examples of serious games are Treefrog Treasure to teach mathematics (Figure 1, left) and Darfur is Dying to raise awareness about the humanitarian crisis in Sudan (Figure 1, right).

The interest of applying games in education has increased in recent years, not only in the education or research communities. Some commercial videogames have also launched their educational versions, to be used by teachers in schools or high schools, such as: a version of SimCity (Electronic Arts Games, 2019) to teach about city management and pollution (Electronic Arts, 2013), a version of Civilization (Firaxis Games, 2016), to teach historical problem solving (Seppala, 2016), an Education Edition of Minecraft (Mojang Studios, 2011) to teach basic coding concepts (Mojang

Studios, 2016), or a version of Portal 2 (Valve, 2011) to teach physics concepts (Valve, 2012).

Figure 1. Screenshot of the serious game Treefrog Treasure (left) retrieved from https://cool-math.co.uk/treefrog-treasure/, and of the serious game Darfur is Dying (right) retrieved from https://www.commonsense.org/education/game/darfur-is-dying.

Despite their multiple advantages, serious games are still rarely applied in education (Kato & Klerk, 2017). One of the reasons for their low adoption is the lack of evidences about the impact they have on players, as no clear evidences are given on how to use games for players' assessment. Therefore, their application is usually limited to a simple complementary or additional activity with no real impact on students' final evaluations (Pereira, De Souza, & De Menezes, 2016).

Some literature reviews have pointed out the potential positive impact of gaming with respect to learning, skill enhancement and engagements, finding that the most frequently occurring outcomes and impacts were knowledge acquisition/content understanding, and affective and motivational outcomes (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012). Besides, prior to assess students by using serious games, the serious game must be evaluated itself. In this regard, there are few approaches to systematically evaluate educational games (Petri & Gresse von Wangenheim, 2017).

In the following subsection, we describe the current assessment techniques of players using serious games, and their possible improvements to provide a more direct measure to assess players directly from their actions in the game.

2.1.1. Players' assessment using serious games

Despite the multiple advantages of applying games in educational contexts, there is a lack of formal or systematic evaluation with games. Few empirical studies have investigated the effectiveness of SGs in learning (Girard, Ecalle, & Magnan, 2013).

The first step before applying serious games in educational contexts is to perform a formal validation of the games, to ensure that they produce the intended effect on players (teaching knowledge, raising awareness, etc.). To determine if serious games have the intended impact on their players, the most common validation technique is called pre-post experiments (Calderón & Ruiz, 2015). This game validation process is as follows (Figure 2):

1. Players complete a questionnaire before playing (*pre-test*) assessing the characteristic the game aims to change (e.g. knowledge, awareness).
2. Players play the serious game, from beginning to end.
3. Players complete a questionnaire after playing (*post-test*), again assessing the characteristic the game aims to change.

The results on the pre-test and the post-test are then compared: if a significant improvement is found between both results, we can say that players have learned, and the game is considered to be effective and, therefore, it is validated. The pre-test and the post-test have been traditionally carried out on paper and, typically, they contain the same questions, or at least questions similar enough so that they can be fairly compared.

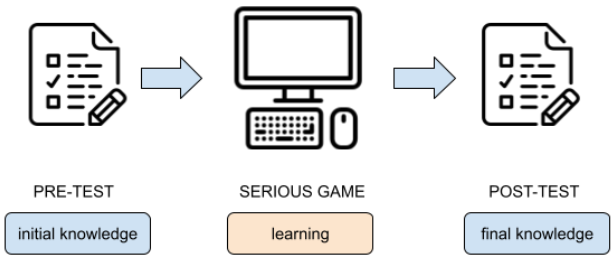

Figure 2. Traditional formal assessment methodology with serious games: pre-post evaluation.

Notice that in Figure 2, and in other places of this document, we simplify the narrative regarding the purposes of serious games by saying that they increase players' knowledge (i.e. games for learning): however, our explanations are equally valid for serious games that aim to cause a different effect on players, such as raise their awareness about some issue, or change their attitudes.

The comparison between pre-test and post-test results provides, for each player, the assessment of their learning and, globally, if results are positive, provides the validation of the serious game. Once the serious game has been formally validated with this approach, questionnaires are also used in deployment. Both pre and post questionnaires are needed to measure the exact learning of players. It is also possible

to assess students only with post-questionnaires, if it is only required to measure final players' knowledge.

The use of external questionnaires to evaluate players has several disadvantages, as they are error prone and increase the total time of the experiment (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012; Frederick-Recascino, Liu, Doherty, Kring, & Liskey, 2013). Questionnaires have, first of all, to be created and validated so that they provide an effective validated measure of the characteristic that they aim to evaluate. Once such questionnaires exist, they have to be prepared and distributed in advance before the gameplay and after the gameplay, therefore, reducing the time left in the session to actually play the game. Finally, questionnaires results have to be digitalized (if they are paper based as they have been traditionally), and analyzed to obtain the evaluation results for all players. Moreover, the complexity and requirements of assessment processes through pre-post questionnaires make them neither scalable nor generalizable outside a controlled experimental setup.

Game-based assessment offer better opportunities than traditional external questionnaires (de Klerk & Kato, 2017), as they provide rich interaction data that can more effectively measure the change on players as it is happening, that is, while they are playing the game. The field of *stealth assessment* captures evidences of gameplay without disrupting players' progress and then compares the evidences gathered in the log files against an evidence model (V. J. Shute & Moore, 2017; V. Shute & Kim, 2014; V. Shute & Ventura, 2013).

Following this line of research, we consider that better informed assessment of players can be obtained with data from game interactions. The rich and varied information that can be gathered from serious games is encapsulated in the field of Game Learning Analytics.

2.2. Game Learning Analytics

Learning Analytics (LA) was originally defined in the first Learning Analytics Knowledge (LAK) conference as the "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Phil Long & Siemens, 2011; Philip Long, Siemens, Gráinne, & Gašević, 2011). LA can therefore contribute to both teaching and learning practice (Gašević, Dawson, & Siemens, 2015).

LA comprises multiple goals and techniques to collect and analyze data from different learning environments. For instance, the reference model for LA (Figure 3) presented by (Chatti, Dyckhoff, Schroeder, & Thüs, 2012) describes the:

- **What?** The variety of educational data that can be gathered from different learning environments to be used for the analysis.
- **Why?** The main goals of LA, including monitoring of students' activities and reporting of results, prediction of knowledge, assessment, adaptation of the learning resources, or reflection about the effectiveness of the learning or teaching practice.
- **How?** The techniques applied including visualizations and data mining.
- **Who?** The different stakeholders that can benefit from the application of LA including students, teachers, institutions, and researchers.

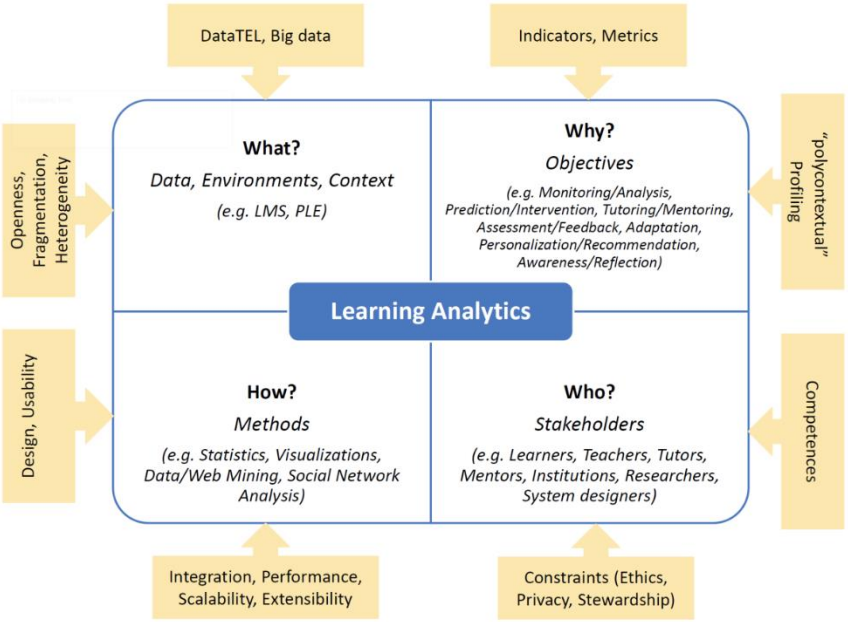

Figure 3. Learning Analytics diagram, retrieved from (Chatti et al., 2012).

The application of LA in education has increased mainly promoted by the large amount of interaction data that can be captured by two systems: Massive Open Online Courses (MOOCs), in which the main focus is to predict student success and dropout (Moreno-Marcos, Alario-Hoyos, Munoz-Merino, & Delgado Kloos, 2018) and Learning Management Systems (LMSs) in educational institutions, also with the focus of early detecting students at risk and take the corresponding actions to prevent their failure in the module (Macfadyen & Dawson, 2010).

The work of (Ruiperez-Valiente, 2020) proposes an implementation process of LA in learning environments (e.g. educational videogames) considering the steps of data collection, data cleaning and feature engineering, and analysis of the data (with exploration, or prediction models) with different goals based on the stakeholder: visual dashboards for teachers, adaptive contents and recommendation systems for students, or educational reports for institutions (Figure 4).

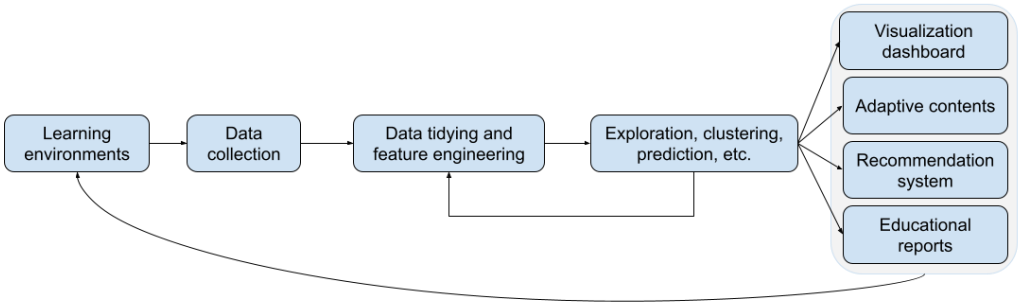

Figure 4. Implementation process of learning analytics, adapted from https://es.slideshare.net/emadridnet/20201113-aplicando-analticas-de-aprendizaje-en-un-juego-serio-de-puzles-geomtricos-jos-a-ruiprez.

The collection and analysis of interaction data is not exclusive to the educational domain. In the field of entertainment games, the analysis of player interaction data is also becoming increasingly widespread. The so-called field of Game Analytics (GA) (Seif El-Nasr, Drachen, & Canossa, 2013) aims to provide data-driven information to support decision-making in the fields of game development and research. With this information, the goal is to better understand players behavior, improve the game design, and ultimately enhance the commercial aspect of the game. Analytics data can improve all stages of the lifecycle of games, including design and development: informing game design based on players' requirements and improving the player experience; discovering potential problems in development and reducing costs; optimizing the game for publishing; improving game user retention and increasing game revenue; and finally, extending the game's life cycle (Su, Backlund, & Engström, 2020).

Moreover, there are some overlapping research areas such as Serious Game Analytics, which focuses on skills and performance improvement (Loh, Sheng, & Ifenthaler, 2015a). In addition, LA can also be applied in serious games, by focusing on the interactions that are meaningful to the learning process (Chaudy, Connolly, & Hainey, 2014). Research applying LA to SGs has focused on learner performance and game design strategies (Liu, Kang, Liu, Zou, & Hodson, 2017). The information gathered can provide feedback to improve and validate the game design: to obtain comprehensive results, authors have stressed that interpretable data should be designed early on, selecting the suitable analysis features (E. Owen & Baker, 2019).

The data that can be collected using LA could be classified as intensive or extensive data (Shoukry, Göbel, & Steinmetz, 2014). Intensive data is obtained when the focus is on a limited number of students, for whom very detail interactions are collected. Extensive data is obtained from a large number of users, when only few data is gathered

about each user results. A combination of both approaches is recommended to complement each other and to avoid missing significant patterns.

The fields of LA and SGs also provide mutual benefits: while LA can improve SGs by providing information that can lead to better game designs or educational results of the game, SGs can also benefit LA by providing an innovative scenario for learning, enriching the opportunities of LA for more accurate and objective assessments (Petrov, Mustafina, Alloghani, Galiullin, & Tan, 2018).

Game Learning Analytics (GLA) is the combination of the educational goals of Learning Analytics with the tools and technologies of Game Analytics (Freire et al., 2016). An implementation of GLA needs to obtain evidences of players' interactions in the game, storing detailed information for later analysis. These analyses could include both real-time analytics to allow to make targeted interventions while students are playing, and later batch analysis, to perform more complex analysis and aggregating results from multiple gameplays. In the conceptual GLA System and Model (Figure 5), the cycle starts when the game sends data to a collector for storage and aggregation, creating the needed reports and visualizations. The information obtained can also be used to assess students. Finally, the analytic system provides feedback to the game to adapt to players' characteristics (Hauge et al., 2014).

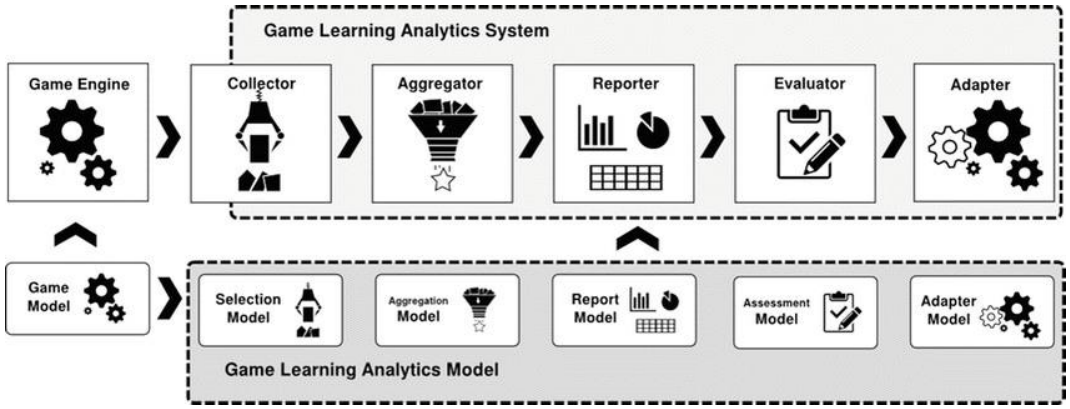

Figure 5. Game Learning Analytics model retrieved from (Hauge et al., 2014).

This way, the results obtained from GLA can be used for multiple purposes including visualizations in dashboards for real-time intervention, offline reporting, adaptation of the game, assessment of players, etc.

The interaction data collected from serious games is very varied and reliant on the game. As the collected interaction data is commonly closely tied to the particularities of the game, it becomes particularly difficult to reuse the interaction data outside the particular environment where it was defined: the combination of data of multiple systems, the integration of tools and even the sharing of interaction data for research

purposes are therefore limited. To simplify the collection, analysis and reporting of results from game interaction data, it is convenient that the collected data format follows a standard. Doing this, it will additionally simplify its integration with other analytics tools that use the same standard.

2.2.1. Data standardization: xAPI-SG

Data standardization is an essential step that can help to simplify the full process of applying game learning analytics data for serious games: to define and simplify the collection of data from serious games, systematize the analysis and report of results, allow the replicability of the process and results, simplify the integration with other environments, and even to share the collected data for other purposes (Kitto, Whitmer, Silvers, & Webb, 2020). Despite the convenience of applying standards, there is a lack of standardization around collecting, analyzing, and managing student learning data from educational games (Keehn & Claggett, 2019).

Currently there are two prominent standards related to the collection of analytical data: IMS Caliper and ADL xAPI. Although they share the same purpose, that is, provide the means to collect and share analytical data between tools, their approaches are different. On the one hand, IMS Caliper defines both a general framework and a set of specific and fixed set of metrics profiles which models specific learning activities interactions. On the other hand, the xAPI specification defines a general framework (API and data model) and the means to let the community define their own profiles. In our work, due to the availability of a specific profile for serious games, as described in detail below, we focus on xAPI. The Experience Application Programming Interface (xAPI, for short) is a data specification created by a community led by the working group Advanced Distributed Learning ADL (ADL, 2012), a part of the Department of Defense of the United States of America. xAPI is based on activity streams (Snell et al., 2011), a standard to represent activities, and aims to provide a standard to communicate information about learners' activities in learning systems. The main concepts of xAPI are verbs, activity types and extensions. Data traces in xAPI (called *statements*) are JSON-based and represent learning activities. Each statement contains three main fields: actor, verb, and object. The actor represents the one who carries out the action, the verb is the action itself, and the object is the item that receives the action. Extensions may be included in the statements to provide further context, results, etc. An example xAPI statement can be seen in Figure 6, representing that a learner has started a new activity.

```
{
    "actor": {
        "mbox": "mailto:john.doe@adlnet.gov",
        "name": "John Doe",
        "objectType": "Agent"
    },
    "verb": {
        "id": "http://adlnet.gov/expapi/verbs/initialized",
        "display": {
            "en-US": "initialized"
        }
    },
    "object": {
        "id": "http://adlnet.gov/expapi/activities/example",
        "definition": {
            "name": {
                "en-US": "Example Activity"
            },
            "description": {
                "en-US": "Example activity description"
            }
        },
        "objectType": "Activity"
    }
}
```


Figure 6. Example xAPI statement representing the event "John Doe initialized the example activity" generated with https://adlnet.github.io/xapi-lab/.

For fields that have specific requirements that go beyond the ones defined in xAPI, Experience API Profiles can be created to provide the means to comply with expertise in that topic area. An xAPI Profile is defined as "the human or machine-readable documentation of application-specific concepts, extensions, and statement templates used when implementing xAPI in a particular context" (ADL, 2017). xAPI Profiles provide specific sets of verbs, activity types and extensions to meet the needs of the topic area. There are several xAPI domain-specific profiles that have been authored (described in http://xapi.vocab.pub/browse/index.html), including profiles for open e-book tracking or healthcare training scenarios.

The xAPI Profile for Serious Games (xAPI-SG) was created to identify and standardize the common interactions that can be tracked in serious games. An interactions model for serious games was created and then validated and published with ADL (Serrano-Laguna, Martínez-Ortiz, et al., 2017) to be the official xAPI Profile for Serious Games, as part of Ángel Serrano's thesis (Serrano Laguna, 2017). The xAPI-SG Profile vocabulary[1] defines a set of **verbs** (*accessed, completed, initialized, interacted, pressed, progressed, released, selected, skipped, unlocked, used*), **activity types** (*area, controller, cutscene, dialog-tree, enemy, item, keyboard, level, menu, mouse, non-player-character, quest, question, screen, serious-game, touchscreen, zone*) and **extensions** (*health, position, progress*) to collect the most common interactions in

[1] http://xapi.e-ucm.es/vocab/seriousgames

serious games. To define their use, the verbs and activity types are related and categorized based on the following higher-level target types:

- **Completables** define something that players can start, progress and complete. The activity types *serious-game*, *level* and *quest* are completable types, and the actions performed with them are *initialized* to indicate the start, *progressed* with the extension *progress* to describe how far the player advances in the current completable and *completed* to signal its ending.
- **Reachables** or accessibles define virtual spaces in the game that players can access or skip. The activity types *screen*, *area*, *zone* and *cutscene* are types of reachables, and the verbs *accessed* and *skipped* are used to track the actions regarding entering those areas, or skipping them, respectively.
- **Alternatives** define decisions that players face in the game. They include the types *question*, *menu* and *dialog-tree*, and the verbs used to track their actions are *selected*, to indicate a choice taken within an alternative, and *unlocked* to indicate that a previously-locked option can now be selected.
- **Targets** define game elements that the player can interact with. They include the types *enemy*, *non-player-character* and *item*. The verbs associated with them are *interacted* for general interactions, and *used* when the player has actually used the target.
- **Devices** define pieces of hardware that the player interacts with to control the game, including *mouse*, *keyboard*, *controller*, and *touchscreen*. The verbs *pressed* and *released* are used to describe interactions with such devices.

Figure 7 displays an example xAPI-SG statement, stating that John Doe (*actor, name*) has selected (*verb*) the incorrect (*result, success*) response Lisbon (*result, response*) in the question (*object, definition, type*) Capital_of_Spain (*object*). Additionally, the extensions in the *result* field collect that the player has a health of 0.34.

```
{
  "actor": {
    "name": "John Doe",
    "mbox": "mailto:johndoe@example.com"
  },
  "verb": {
    "id": "https://w3id.org/xapi/adb/verbs/selected",
    "display": { "en-US": "selected" }
  },
  "object": {
    "id": "http://rage.e-ucm.com/activities/Countrix/questions/Capital_of_Spain",
    "definition": {
      "type:" "http://adlnet.gov/expapi/activities/question"
    }
  },
  "result": {
    "response": "Lisbon",
    "success": false,
    "extensions": {
      "https://w3id.org/xapi/seriousgames/extensions/health": 0.34
    }
  }
}
```


Figure 7. xAPI-SG sample statement generated when "John Doe selected a false response in a question", retrieved from (Serrano-Laguna, Martínez-Ortiz, et al., 2017).

The use of the xAPI-SG Profile can simplify the collection of interaction data from serious games. The collected data can then be used for multiple purposes, such as the assessment of players. For that and other purposes, multiple techniques can be applied to the GLA data gathered, including data mining techniques.

2.3. Data mining techniques

The potentially large amount of GLA data collected from players interactions in serious games can be analyzed with multiple techniques, including data mining techniques. Data mining (DM) is a term defined as the "process of discovering interesting patterns and knowledge from large amounts of data" (Han, Kamber, & Pei, 2012). This technique is commonly a step in a large process of knowledge discovery, that involves the following:

1. **Data cleaning**: removes noise and inconsistent data.
2. **Data integration**: combines multiple data sources.
3. **Data selection**: choses the relevant data for analysis.
4. **Data transformation**: performs aggregations and other operations to shape data in the appropriate forms.
5. **Data mining**: applies intelligent methods to extract data patterns.
6. **Pattern evaluation**: identifies the interesting patterns and knowledge.
7. **Knowledge presentation**: presents the results through visualization or representation techniques.

Prior to creating any DM models, data preprocessing is essential to clean the dataset finding possible missing or incorrect values, integrate all the data from the different data sources, remove any redundant data and selecting the data variables, and perform the required transformation on the data variables. After performing the data mining models and evaluating the results, the last step of the DM process commonly includes the creation of visualizations to report information about the gathered data; in education, visualizations are usually put together in teacher or learner dashboards, providing an overview of the actions taken.

The DM techniques mainly focus on the problems of supervised and unsupervised learning. Supervised learning comprises the techniques that classify the new input data from the labeled training data points. Unsupervised learning, however, comprises the techniques to create clusters of the dataset, where input examples are not labeled in classes. These learning techniques can be applied, for instance, to create prediction models of players' learning, in the case of supervised learning, or to create different player profiles based on their actions, with unsupervised techniques.

In this work, we focus on supervised techniques: prediction models that can be created for both classification and regression problems, depending on whether the target variable is categorical (or binary), or linear, respectively. Classification models include decision trees, Bayesian classification, or support vector machines (SVM), while regression models include linear regression or regression trees. Some models like neural networks or k-nearest neighbors (k-NN) can be adapted to both classification and regression problems. Ensemble methods combine multiple models to improve performance: for instance, for classification, an ensemble classification model is made up of a combination of different classifiers, each of one vote and the ensemble final classification is based on the combination of the individual votes. Random forests or Ada boost models are examples of ensemble models.

For the work carried out in this thesis, we selected a variety of prediction models based on the ones commonly reported in similar works in the literature (as later described in the results of the literature review) and including both traditional models (e.g. regression and tree-based models), as well as some of the more complex and promising models (e.g. neural networks and ensemble methods) to try to improve the results. As a brief summary, the prediction models included in this thesis for classification problems, as defined in references like (Agarwal, 2014; Irizarry, 2019), are:

- **Decision trees** are flowchart-like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf or terminal node holds class label.

- **Logistic regression** is a simple linear model, that converts probabilities into labels where each outcome is assigned to one of the two numeric values of 0 and 1. The model can be extended to model several classes.

For regression problems, the prediction models considered in this thesis include:

- **Regression trees** are built exactly like decision tress, but the predictive value in each leaf node is not a class, but an average value of the training observations that also fall in that leaf.
- **Linear regression** is a simple linear model that finds the "best" line to fit two attributes (or variables) so that one can be used to predict the other, that is, the data are modeled to fit a straight line. This technique can be extended to more than two attributes, in what is called multiple linear regression.
- **Support Vector Machines (SVM)** use a nonlinear mapping to transform the original data into a higher dimension, in which it is possible to create the "best" hyperplane to separate the data. SVM has its corresponding version for regression, SVR, which can use both linear and non-linear kernels.
- **Bayesian ridge regression** is a version of a Bayesian regression model that includes regularization parameters.
- The **k-nearest-neighbor (k-NN)** algorithm represents data in a multidimensional space, and bases the prediction for each input data on the values of the data that are closest in that space to the input data (the nearest neighbors).
- **Neural networks** are a set of connections between input and output units, with an associated weight for each connection. During the learning phase, the network adjusts the weights so that it performs the best prediction on the input.
- **Random forests** are an ensemble method that combines multiple single regression or decision trees, whose predictions are combined in each step to obtain a global prediction.
- **AdaBoost** (adaptive boosting) is a type of boosting algorithm that combines multiple methods and weighs their predictions, changing the weight of each algorithm based on their accuracy in the previous iterations.
- **Gradient boosting** is a type of boosting algorithm that ensembles multiple prediction models (typically trees).

Finally, to evaluate the predictive efficacy of the models, it is necessary to test them. For that purpose, we performed cross validation (CV), a technique that divides the training data into multiple groups and performs as many iterations as groups to ensure that all data points participate in both training and testing steps of the prediction

models. Different metrics can be applied to test the models' effectiveness. For the classification models considered in this thesis, we selected the metrics of precision and recall, to provide an indication of how *successful* the models were, as well as misclassification rate (MR) to provide a measure of *error*. These metrics are based on the defined values of the confusion matrix (Table 1) regarding the relation between positive and negative values, and their predicted ones.

Table 1. Confusion matrix for classification algorithms.

	Positive Value	**Negative Value**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

In particular, the three metrics that we report for the classification algorithms used, precision, recall and misclassification rate (MR), are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$MR = \frac{FP + FN}{n}$$

For the regression algorithms, we chose the metric of Mean Absolute Error (MAE), as it is an interpretable measure of the error of the models. The MAE is defined as:

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Where the absolute error $|e_i| = |x_i - y_i|$, being x_i the true value and y_i the predicted value.

Data mining techniques have been extensively applied in many fields, including education. Educational Data Mining (EDM) is defined as a discipline "concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (R. Baker & Yacef, 2009). EDM is related to LA as it applies techniques from the fields of statistics, machine learning, and data mining to analyze data gathered during teaching and learning activities (Bienkowski, Feng, & Means, 2012; ElAtia, Ipperciel, & Zaïane, 2016). The works on EDM have focused on the analysis and visualization of data, providing feedback to support instructors, providing

recommendations for students, predicting student performance, student modeling, detecting student behaviors, studying the relationships between students, grouping students, and creating and planning the course (Romero & Ventura, 2010).

Figure 8 describes the full process of EDM (Vahdat et al., 2015) that includes the collection of data, and its preprocessing, the data analysis to obtain metrics and the postprocessing to report feedback and create interventions to optimize the learning process. Notice the similarities between this process and the GLA system depicted in Figure 5, including the collection of data from the learner/player, the analysis and reports, and the final feedback step to close the process.

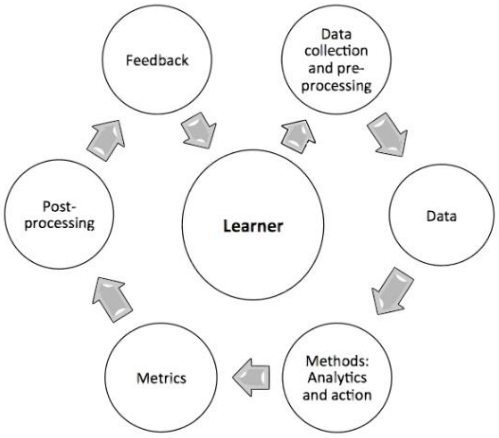

Figure 8. Educational Data Mining (EDM) process, retrieved from (Vahdat et al., 2015).

2.4. Applications of data mining to game learning analytics data

In the previous sections we have described, separately, the three main topics of our thesis: serious games, game learning analytics, and data mining techniques. In this section, we focus on the combination of the three topics: the application of data mining techniques to game learning analytics data from serious games. As we did not find many studies that considered all three aspects, we carried out a specific systematic literature review, which constitutes one of the contributions of the thesis, as described below.

Data science and artificial intelligence (AI) have been applied in games to study game-playing, content generation and player modeling (Yannakakis & Togelius, 2018). The applications of learning analytics on serious games for assessment has also been studied, and authors have found out how SGs had a positive impact on learning and highlighted the importance of the game design (Liu et al., 2017). Despite these and other works previously discussed, we did not find literature reviews or major revision studies that

combined the three areas of our interest: serious games, game learning analytics data, and data mining techniques.

To fill this gap identified in the literature, we carried out a specific systematic literature review focusing on the specific data science techniques used to game learning analytics from serious games. Specifically, we analyzed the purposes for which data science had been applied to GLA data from SGs, the data science techniques applied, the stakeholders that were the targets of such analysis, and the results obtained in the studies (Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2019). We additionally gathered the information about the serious games used in the studies, the interaction data captured, and the participants included in the experiments. The rest of the section describes the key aspects of the related work found out as a result of the literature review, with emphasis on the results obtained in the studies and the discussion and conclusions pointed out in such works. Further details about the results of the literature review, and the overview of the main results, are included in section 4.1 of this document, as part of the results and contributions of the thesis.

In the studies reviewed, the main purposes of applying data science to GLA data from SGs were player assessment and performance predictions, the study of players' in-game behaviors and the validation of the games design. The data science algorithms and techniques used in the studies can be grouped into three main categories: regarding supervised algorithms, a majority of studies used linear and logistic regression, or regression and decision trees; the unsupervised algorithms most commonly included in the studies were correlation and clustering; and performance metrics were usually included in visualizations. The applications included in the studies were mainly targeted at serious game designers and developers, researchers, or teachers and educators. The focus of the serious games used was to teach (in particular, mathematics and science), to participants of primary and secondary school. The sample sizes included in most studies were low (less than 100 participants). The interaction data captured mainly focused on completion, scores and interactions in general; while the data format in which interactions were collected was majority not stated.

The results obtained in the application of data science techniques to GLA data were varied and closely related to the purpose of application. To encapsulate the results, we defined three main groups: (1) studies focusing on players' assessment and learning predictions; (2) studies focusing on serious game design and implementations; and (3) studies that propose frameworks to apply GLA in specific contexts.

In the first group of studies identified in the review, studies focused on assessment of players and learning predictions. These studies highlighted how GLA data can accurately predict and measure games' impact (Kosmas, Ioannou, & Retalis, 2018; Mavridis, Katmada, & Tsiatsos, 2017): being useful at real-time and after the interventions are completed (Wiemeyer, Kickmeier-Rust, & Steiner, 2016), and for all stakeholders (Alonso-Fernández, Pérez-Colado, Freire, Martínez-Ortiz, & Fernández-Manjón, 2019). However, authors have pointed out that most data are still captured after the game (Smith, Blackmore, & Nesbitt, 2015) and that specific game learning analytics (Freire et al., 2016), or so-called serious games analytics (Loh, Sheng, & Ifenthaler, 2015b) are required for more precise information. Regarding learning predictors, the achievement system built into games may not be the most informative indicator of learning (Heeter, Lee, Medler, & Magerko, 2013); instead, predictions of player success should be based on log data (R. S. Baker, Clarke-Midura, & Ocumpaugh, 2016; Rowe, Asbell-clarke, & Baker, 2015). The best learning predictors are based on the analysis of the player's exploration strategies (Horn et al., 2016; Kang, Liu, & Qu, 2017; Käser, Hallinen, & Schwartz, 2017; V. E. Owen, Anton, & Baker, 2016; Smith, Hickmott, Southgate, Bille, & Stephens, 2016), or on player failures (Halverson & Owen, 2014) and behaviors (Hernández-Lara, Perera-Lluna, & Serradell-López, 2019; Ketamo, 2013; Mayer, van Dierendonck, van Ruijven, & Wenzler, 2014; Rowe et al., 2017; Tellioglu, Xie, Rohrer, & Prince, 2014; Z. Xu & Woodruff, 2017).

Authors also provided some recommendations to improve the learning predictions: perform feature engineering (V. E. Owen & Baker, 2018), include the domain structure and the weights of competencies (Kickmeier-Rust, 2018), and perform exploratory data analysis (DiCerbo et al., 2015) and dynamical analysis (Snow, Allen, & McNamara, 2015) to uncover unexpected patterns. Assessment can be further improved combining generic game trace variables (Steiner, Kickmeier-Rus, & Albert, 2015) or basic sets of traces (Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2014).

Many of the studies also analyzed how performance was related to players' characteristics: creating clusters of players performance based on their actions (Chung, 2015; Cutumisu, Blair, Chin, & Schwartz, 2017; Forsyth et al., 2012; Freitas & Gibson, 2014; Lazo, Anareta, Duremdes, & Red, 2018; Martin et al., 2015; Martinez-Garza & Clark, 2017; Polyak, von Davier, & Peterschmidt, 2017; Sharples & Domingue, 2016; Slimani, Elouaai, Elaachak, Yedri, & Bouhorma, 2018); or differentiating experts from novice users (Loh & Sheng, 2014, 2015a, 2015b). Once students are classified in a performance group, scores can be inferred adding time or action sequences (Gibson & Clarke-Midura, 2015). Learners' characteristics such as age and gender (Wallner &

Kriglstein, 2015), background (Jaccard, Hulaas, & Dumont, 2017), or exploration strategies (Martin et al., 2013) also influence their learning behaviors (Liu, Lee, Kang, & Liu, 2016). Modelling students is essential to effectively adapt learning (Koedinger, McLaughlin, & Stamper, 2012; Liu, Kang, Lee, Winzeler, & Liu, 2015; Sabourin, Shores, Mott, & Lester, 2013). GLA data can also be used to track students' progress (Gweon et al., 2015), assess persistence (Dicerbo, 2013), or detect engagement (Ghergulescu & Muntean, 2016). The information gathered can also be presented at real-time to teachers and students (Elaachak, Belahbibe, & Bouhorma, 2015) or parents (Ketamo, 2015; Roberts, Chung, & Parks, 2016).

The second group of studies focused on improving serious game design and implementation. First, their results proved that GLA data can be used to validate the serious game design (Cano, Fernández-Manjón, & García-Tejedor, 2018; Cheng, Rosenheck, Lin, & Klopfer, 2017; Harpstead, MacLellan, Aleven, & Myers, 2015; Ninaus, Kiili, Siegler, & Moeller, 2017; Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2012; Tlili, Essalmi, Jemni, & Kinshuk, 2016). Authors have stressed how assessment should be early integrated in serious games development and design (Ke, Shute, Clark, & Erlebacher, 2019; Ke & Shute, 2015), starting from an early definition of the game traces to be collected (Serrano-Laguna, Manero, Freire, & Fernández-Manjón, 2017; Tlili et al., 2016). To improve the design of games for assessment, studies have pointed out that assessment models should be reliable, providing meaningful educational information (Steiner et al., 2015). For that, it is required to explore how design decisions affect the learning outcomes (Plass et al., 2013) and include adaptivity (Streicher & Smeddinck, 2016). Learning has also been investigated in relation to some serious games' characteristics: adaptive difficulty (Hicks et al., 2016; Käser et al., 2013; Martinez-Garza & Clark, 2017); engagement and motivation (Pareto, 2014; Stamper et al., 2012; Tlili et al., 2016); and feedback and interventions during play (DeFalco et al., 2018; McCarthy, Johnson, Likens, Martin, & McNamara, 2017).

The final group of studies included the proposal of frameworks to simplify serious game design in specific contexts: game analytics frameworks for people with intellectual disabilities (García-Tejedor, Cano, & Fernández-Manjón, 2016; Nguyen, Gardner, & Sheridan, 2018), a game-based assessment model (Halverson & Owen, 2014), a framework to integrate design of event-stream features for analysis (V. E. Owen & Baker, 2018), a framework to support tracking and analysis of learners in-game activities (Hauge et al., 2014), a framework to help designers model experts' solving process almost automatically (Muratet, Yessad, & Carron, 2016), an

interoperable adaptivity framework (Streicher & Roller, 2017), a framework for internet-scale experiments to inform and be informed by classroom and lab experiments (Stamper et al., 2012), an open-source SGs framework for sustainability (Y. Xu, Johnson, Lee, Moore, & Brewer, 2014) and a framework for a mobile game application for adults with cystic fibrosis (Vagg et al., 2018).

From the review we can highlight several conclusions. Most studies focused on assessment and learners' behaviors. Games are indeed a useful tool for purposes beyond entertainment, so now the interest focuses on analyzing interaction data to measure how much impact serious games have on players (mainly focusing on learning), and how that impact relates to players' in-game behaviors. Studies used visualization, supervised and unsupervised techniques, mainly linear models, correlations, and cluster techniques. Newer and more complex and powerful techniques, like neural networks, are experiencing an important surge in popularity, but they did not appear that frequently in the reviewed studies. One possible explanation is that further evidence is needed on how to widely and reliably apply these new complex techniques, as well as to explain the results obtained, an open debate about explainable AI (xAI) (Adadi & Berrada, 2018).

The main stakeholders considered are game designers/developers, and researchers, followed by teachers/educators. This suggests that the analysis of data from games is used for several purposes including research, improving or validating game design, and providing information when applying games in educational scenarios. Still, students are always indirect recipients of the results, as the research, improvement and adaptivity of games and assessment techniques will make the use of games more effective and efficient for the ones who play the games, that is, the students/learners.

Most of the games used in the studies aim to teach science-related topics, in particular mathematics. This result shows the intention to benefit from games' advantages to improve learning in a subject typically difficult for young students. It also and aligns with previous research which identified mathematics and science as the main areas for games targeted at primary education (Hainey, Connolly, Boyle, Wilson, & Razak, 2016).

Sample sizes used in the studies are, in general, quite low (less than 100 participants). This may restrict the significance and generalization of their results, as well as the application of more complex algorithms (e.g. deep neural networks), which require larger amounts of data points. The low sample size used in experiments is an important issue, pointed out by authors (Petri & Gresse von Wangenheim, 2017).

Data collected from students' interactions included mainly completion times, actions or interactions in general, and scores. All these data can be collected from most games, but provide basic information that does not take full advantage of the rich interactions produced in games, as described in works on game analytics in entertainment games (Seif El-Nasr et al., 2013). The data to be collected should be identified at early stages of the game development, to ensure that it provides information with educational value. Most papers did not report the format in which they collected the data, so we cannot know if they were using a standard or relying on their own data-formats. The latest scenario is less desirable, as it restricts the open sharing of the data for other purposes and requires an extra effort to replicate results with other techniques (Serrano-Laguna, Martínez-Ortiz, et al., 2017). We have not found reports of any open data set of game analytics data or learning analytics data from serious games; this hinders research in this area, as testing out new data science techniques requires not only choosing the techniques themselves, but also developing a serious game and performing the experiments to collect its interaction data.

The analysis of GLA data from serious games has yielded, as expected, wide and varied results. We can, however, extract some general findings from the conclusions and discussions of the studies analyzed:

- **Predicting games impact with GLA data:** raw data can be used to accurately predict impact (e.g. learning), including simple values from interactions (e.g. completion times, scores) but also more complex information such as kind of failures or exploration strategies. Adding information of the context is also recommended, as it can improve the models' accuracy. Also, the choice of data to analyze should ideally be taken during game design, to ensure that as much educationally relevant data as possible is actually captured.
- **Importance of student profiling:** performance appears to be highly related to students' characteristics and behaviors, so it is recommended to create students' profiles or clusters to improve learning, including targeted feedback and adaptive learning experiences. The need to fit users' needs has also led authors to propose user-specific frameworks (e.g. for users with intellectual disabilities).
- **Designing serious games for assessment:** assessment needs to be formally and reliably integrated in the development phase of SGs to provide meaningful educational information. This should not damage costs or entertainment, as games need to maintain engaging and motivation features, while controlling for an adequate difficulty. GLA data can then be used to validate the game design and assessment.

From the conclusions obtained in the studies reviewed, we can further point out that, although plenty of studies have started to investigate the opportunities that GLA data offers to improve the full lifecycle of serious games (from design and development, to deployment in real scenarios, and the assessment of players), there is still a gap in the literature regarding the systematization and generalization of players' assessment. That is, the studies that investigate this (including *stealth assessment* techniques) are very tied to the game they are applied to, turning players' assessment into a process performed ad-hoc for each specific case. More research is needed to try to generalize these data-based approaches, providing tools to systematize the steps required to effectively and accurately assess students based on their game interaction data. Additionally, such studies should also consider large-enough sample sizes, and more so if they involve the application of complex data mining techniques (e.g. neural networks), to ensure that there are enough evidences to support their results and replicate them in different scenarios.

Our evidence-based approach, presented in the following chapters, aims to contribute in such identified areas: creating some easy-to-follow steps to assess players based on their interaction data with a serious game, using standards to support the process, and generalizing each step as much as possible so the process can be applied to a wide range of serious games (at least, to serious games with similar features, like narrative-based games), providing tools to support the process.

In conclusion, serious games have a great potential to cause a positive effect on players. Combining the educational purposes of LA, and the techniques of GA, the field of game learning analytics comprises the collection and analysis of interaction data from serious games, that can provide a richer insight into players' actions, inform about their progress and actions, and provide feedback to improve the serious game and the learning process. The information obtained from the analysis of GLA data can additionally be used for players' assessment, improving the commonly method of assessment based on external paper-based questionnaires. This traditional technique does not base the assessment on data collected while learning is happening, that is, while players are playing the serious game. With the GLA evidences collected, new opportunities for assessment arise, as they can provide a finer-grained information on which to assess players.

The GLA data can further be analyzed with data mining techniques, applying lessons from the field of EDM to serious games. Although some works have started to explore this idea, there is a lack of standardization and systematization to create evidence-based assessment of players based on their interactions with the serious game. To try

to advance in this identified gap, we propose the goals of this thesis, described in detail in the following section, to improve the current assessment methods of players using serious games, combining the potential of the information gathered from game learning analytics data, with the richer analysis obtained from data mining techniques.

Chapter 3. Goals of the thesis

This chapter presents the general research goals of the thesis, and how they were divided into specific goals to be achieved during the research; together with the research methodology followed and how the steps in the research process were carried out to achieve these goals.

3.1. Research goals

The main goal of the thesis is to simplify the current method of assessment of players using serious games, taking advantage of the richer value provided by the game learning analytics data created by players' interactions during the gameplay. The current assessment methodology is based on external, frequently paper-based, questionnaires. We aim to avoid the use of these external questionnaires after the formal game validation phase and, instead, collect game learning analytics to obtain richer information about students' gameplays and analyze it with data mining techniques to predict players' performance.

The large-scale goal is that the simplification of players' assessment will, in turn, provide teachers and educators clearer evidences of the impact of playing serious games. If the assessment is then simplified, we consider that this will also foster the application of serious games in a broader set of actual educational scenarios, as clearer evidences will be provided of their impact on players. This can increase the adoption of serious games as providing teachers and institutions with evidences of the effectiveness of using serious games will make them a more attractive choice.

To reach these high-level goals, we proposed the following specific milestones:

1. Analyze the current assessment techniques of players using serious games, the application of Game Learning Analytics data in the context of serious games, with particular focus on their application for assessment, and the application of data mining techniques to GLA data for assessment of players using serious games.
2. Propose an application of data mining techniques to the GLA data collected from serious games to simplify assessment of players using serious games, relying as much as possible on the game interaction data collected.
3. Verify the suitability of the proposal in actual experiments, collecting player interaction data from serious games, and applying data mining techniques to

the GLA data collected to create predictions that can effectively assess students, without requiring the use of pre-post questionnaires.

4. Iterate, as needed, the previous step to improve the evidence-based assessment process of players using serious games, analyzing the steps to move towards their generalization.
5. Create a final version of the evidence-based process to assess players using serious games based solely on the application of data mining techniques to the GLA data collected from their game interactions, and providing the required tools and standards to support the process.

3.2. Research process

The process followed in the thesis was based on the design science research methodology (DSRM) (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), that consists of the following steps (Figure 9):

1. Identify and define the problem
    - Show the importance and motivation
2. Define the objectives of a solution
3. Design and develop a solution
4. Demonstrate the solution in a suitable context to solve the problem
5. Evaluate the effectivity and efficiency of the solution
    - Iterate back to design
6. Communicate the solution
    - Publications

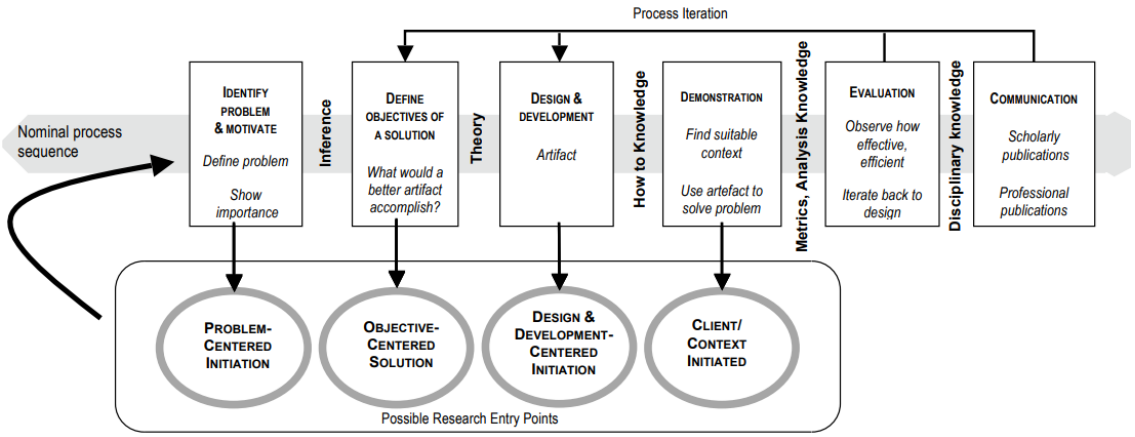

Figure 9. Design Science Research Methodology (DSRM), retrieved from (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

The DSRM was adapted to our context of application and based on our specific goals: to identify the problem, we started by studying the field to find areas for improvement

and issues that other authors had identified that required further research. Once the review was finished and the problem clearly identified, we proposed a first solution and we developed our proposal in two iterations. First, we prototyped the ideas and tested them by carrying out experiments in real settings to collect actual data that can serve us to test our approach, verify its suitability, refine it and improve it. Second, we synthesized the knowledge acquired in the experiments carried out to obtain and distill our final approach, with all the required steps, based on standards and developing the required easy-to-use tools to support the process. All the results were suitability communicated with publications in journals and conferences.

The research process conducted in the thesis was then as follows:

- Review the literature regarding the application of game learning analytics data and data mining techniques in the field of serious games.
- Explore in a first case study the validity of a new players' assessment approach using game learning analytics data and prediction models.
- Conduct a second case study to verify the approach with another serious game with different goals, and different prediction models.
- Create an assessment process based on game learning analytics data and data mining models, as general as possible, and describing each step of the process in detail. The process is based on the conclusions obtained in the two case studies; the use of standards like the xAPI-SG Profile to collect the interaction data from serious games; and includes the development of any required tools to support the process, like the tool T-MON to support a first exploration of the xAPI-SG interaction data collected to refine the selection of GLA variables.
- Publish all results in high ranking journals and in conferences.

Additionally, we had the previous experience of the work carried out in the European Projects H2020 RAGE and, particularly, BEACONING in which we had worked in the full lifecycle of serious games deployed in real large-scale scenarios: from their design and development, to the collection and analysis of the game learning analytics collected from players' actions. Finally, the research stay carried out at Florida State University allowed us to contrast our approach with the work of the research group of Professor Valerie Shute, leader in the field of *stealth assessment*.

Chapter 4. Results and discussion

This chapter presents the results obtained in the thesis including: the systematic literature review about the application of data mining techniques to game learning analytics data from serious games; and the two case studies conducted including their main characteristics and the main analytical insights obtained from them. The final result is the evidence-based assessment process distilled from these case studies, describing the game validation process including the definition of the interaction data to be collected, the use of the standard data format xAPI-SG, the feature extraction process into game learning analytics variables, and the tool to support that selection, T-MON, and the prediction models to simplify players' assessment in games deployment. The chapter finalizes with a discussion of the results obtained, and points out the limitations of our work.

4.1. Study of the domain

The study of the domain was conducted through a systematic literature review of the applications of data science techniques to game learning analytics data from serious games. For that review, we established the following research questions (RQ):

RQ1. What are the purposes for which data science has been applied to game analytics data and/or learning analytics data from serious games?

RQ2. What data science algorithms or techniques have been applied to game analytics data and/or learning analytics data from serious games?

RQ3. What stakeholders are the intended recipients of the analysis results?

RQ4. What results and conclusions have been drawn from these applications?

We additionally included in the review the following information from the studies:

- The main purpose of the games (e.g. teaching, change behavior) and their domain (e.g. biology, math).
- The sample size of the studies, and the educational level of their participants.
- The general characteristics of the in-game interaction data collected, and the data format used.

After defining the suitable search terms and databases, we carried out the selection process depicted in Figure 10, obtaining a final set of 87 studies included in the review.

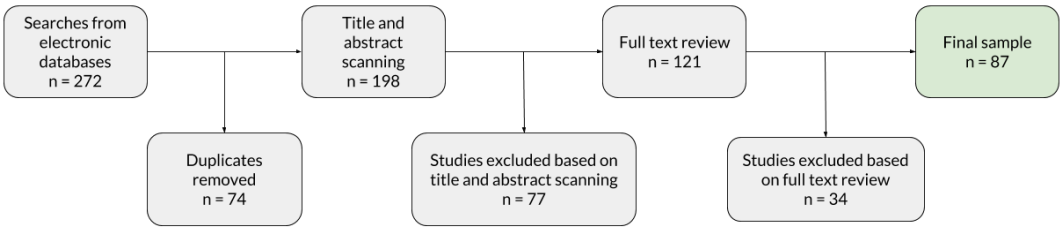

Figure 10. Process carried out to select the publications included in the systematic literature review.

Regarding the applications of data science to game learning analytics data from serious games (RQ1), the main purposes include: assessment of learning, or prediction of players' performance; the study of players' in-game behaviors; the validation of the game design; the profiling of students in different categories based on their actions and characteristics; the study of different in-game interventions on players' performance; and the proposal of different frameworks of application for particular scenarios (Table 2).

These results confirm the interest pointed out by other authors regarding the use of game learning analytics data to assess learning or predict players' performance with serious games, with an additional high interest in studying how players behave in game.

Table 2. Main purposes of data science applications to game learning analytics data from serious games.

Purpose category	Definition of purpose	Example studies	Number of studies
Assessment	Assess learning, predict performance	(R. S. Baker et al., 2016; Ke & Shute, 2015)	32
In-game behaviors	Study in-game players behaviors (e.g. persistence, engagement)	(Dicerbo, 2013; Kang et al., 2017)	27
Game design	Validate game design	(Cano et al., 2018; Tlili et al., 2016)	16
Student profiling	Stablish categories of players profiles, differentiate players characteristics	(Denden, Tlili, Essalmi, & Jemni, 2018; Loh & Sheng, 2014)	7
Interventions	Study effect of in-game interventions (e.g. feedback messages, notification of performance)	(DeFalco et al., 2018; McCarthy et al., 2017)	4
Framework proposals	Propose a framework for specific contexts	(Halverson & Owen, 2014; Nguyen et al., 2018)	10

The data science algorithms and techniques used in the reviewed studies (RQ2) can be grouped into three main categories (Table 3):

- Supervised algorithms: linear and logistic regression, regression and decision trees, support vector machines, Bayesian networks, neural networks, naive Bayes, and Bayesian knowledge tracing.
- Unsupervised algorithms: correlation, clustering, factor analysis.
- Visualization techniques: display of gameplay pathways, performance metrics, learning curves, heatmaps of interactions, use of in-game tools (frequency or duration).

Some of the studies used multiple of the previous techniques. From these results, we noticed that most studies were using simple traditional algorithms (e.g. linear regression, clustering), while not so many studies were using more complex techniques (e.g. neural networks).

Table 3. Data science techniques applied to game learning analytics data from serious games.

Data science technique	Number of papers using the technique
Supervised models	**31**
Linear/logistic regression	18
Regression/decision trees	7
Bayesian networks	6
Neural networks	4
Naïve Bayes	3
Bayesian knowledge tracing	3
Support vector machines	2
Unsupervised models	**35**
Correlation	17
Clustering	16
Factor analysis	2
Visualization	**36**
Performance metrics	15
Gameplay pathways	7
Use of in-game tools	5
Learning curves	4
Heatmaps of interactions	2

Regarding the main stakeholders targeted by the studies (RQ3), serious game designers and developers were the main target (39 studies), closely followed by researchers, or studies carried out with research purposes (37 studies). Many studies (25) also focused on teachers and educators. Finally, some studies mainly focused on students and learners (8) and few mentioned parents (2).

We further analyzed the information that the studies provided regarding the serious games used, the participants included in the experiments, and the interaction data collected from their actions with the serious games. Regarding the serious games used, most of the serious games applied in the studies focused on teaching some knowledge

to the players (55 studies). Most games aimed to teach mathematics (20 studies) or science (10 studies). Aligning with those results, the educational level of the participants was mainly primary, secondary school students, and undergraduates. The sample sizes of the studies were limited: around a third of studies used fewer than 100 participants, and around another third used fewer than 1000 participants, although most of such studies had a sample size closer to few hundreds of players. Finally, the interaction data collected from serious games mainly included completion times (30 studies), general interactions in the game (28) and scores (14). Most studies did not report the format on which they collected the interaction data.

The studies reviewed provided diverse results that allowed us to draw the following general and specific conclusions:

- Regarding players' assessment and student profiling
    - GLA data can accurately predict serious games impact, for instance, predictions can be created based on players' exploration strategies, failures or interactions between players in collaborative games. To improve the predictions, it is recommended to perform feature engineering and combine basis set of traces and variables. GLA information can be used both at real-time and after the intervention, and for all stakeholders.
    - Learning performance is related to players' characteristics, therefore, players can be effectively be clustered into performance groups based on their actions. Understanding their learning characteristics is essential to better predict learning and improve or adapt the learning process.
    - Additionally, further information can be extracted from GLA data to study additional students' characteristics, or to provide real-time information to players, students, and even parents.
- Regarding serious games design
    - GLA data can effectively validate serious games design and mechanics.
    - Assessment can and should be integrated in early stages of serious game design and development, and it should be transparent and reliable, based on models that are valid, easy to use, and provide meaningful educational information. The data to be collected should be specified early on so that meaningful GLA information can be obtained from it.
    - Serious games characteristics, such as difficulty, engagement and motivation, and feedback and interventions during the gameplay, should also be considered as they affect learning.

- Specific frameworks have been proposed to apply GLA in different scenarios simplifying the multiple tasks in serious games design.

In summary, despite the diversity of the studies, we found some notable common points: the main purpose when analyzing data from serious games is learning prediction or assessment, most commonly with linear prediction models, simple correlation or cluster techniques, or visually displaying performance information. Learning predictions obtained are quite accurate and may be improved with some of the pointed-out recommendations (feature engineering, combining multiple traces). The importance of student profiling as well as recommendations for integrating assessment into early phases of game design and development also stand out among the conclusions of the studies.

It also stands out that future research should consider large-enough sample sizes to ensure significant conclusions, and to decide in advance which data is to be collected from the games. In this sense, as a baseline, typical data such as completion times, interactions, or scores can and should be included; but research can benefit from moving on to more complex data extracted from in-game interactions. Regarding algorithms, classical techniques should be compared with new more complex ones (e.g. neural networks), to determine which ones draw the best results in each case. Finally, authors have pointed out a clear need for specific game learning analytics, where the use of standards to collect GLA data is desirable, as it allows the creation of open data sets in standard formats, such as xAPI, for research purposes, and simplifies results reproducibility and improvement, as well as testing of new techniques and integration of analytics as a module of larger systems.

The process, results and details about this systematic literature review have been published in (Alonso-Fernández, Calvo-Morata, et al., 2019), a publication that is included as part of this thesis. For the full text and details of the publication, see subsection 6.1.1.

4.2. First case study

In this section, we describe the first case study carried out to test the approach to assess players automatically from their interactions with a serious game. In this case, we used a game to teach knowledge about first aid techniques. Besides verifying the suitability of our approach, in this first study, we were additionally interested in verifying if students' previous knowledge (given in their pre-questionnaire results) was essential to accurately predict their knowledge after playing.

The following subsections describe the serious game used (4.2.1), the interaction data captured from it (4.2.2), the analysis on that data to create GLA variables (4.2.3), the prediction models used to assess players and the results obtained (4.2.4), and the discussion and conclusions of the case study (4.2.5).

4.2.1. The game: *First Aid Game*

The *First Aid Game* (Marchiori et al., 2012) is a game-like simulation with narrative structure that aims to teach basic life-support maneuvers in the situations of chest pain, unconsciousness and choking. The game is targeted for players between 12-16 years old. The three medical situations are depicted as different game levels. In each level or scenario, players can interact with several game elements: the main character, suffering from the medical condition depicted in that level, or a mobile phone that they can use to call the simulated emergency services. In each of the levels, different multiple-choice situations (second screenshot of Figure 11) are presented for the player to choose the course of action between a set of visual or textual options. These situations include the specific first aid knowledge to be learnt through the game (e.g. Heimlich maneuver to avoid chocking). Players learn if their decisions are appropriate or not: when choosing an incorrect answer, either the game reports the critical error and its consequences, and lets them try again until they choose the correct answer, either the game allows you to continue to later discover the consequences (and it is reflected in the final score). The game includes random elements to improve reflection and replayability (e.g. availability of a semi-automatic external defibrillator that players can also use). The three levels can be replayed as many times as wanted during the allotted time. After each level is completed, a score is provided to indicate players whether their actions were mostly correct or not (first screenshot of Figure 11). The presented score does not directly measure players' knowledge but challenges them to replay levels where they made many mistakes.

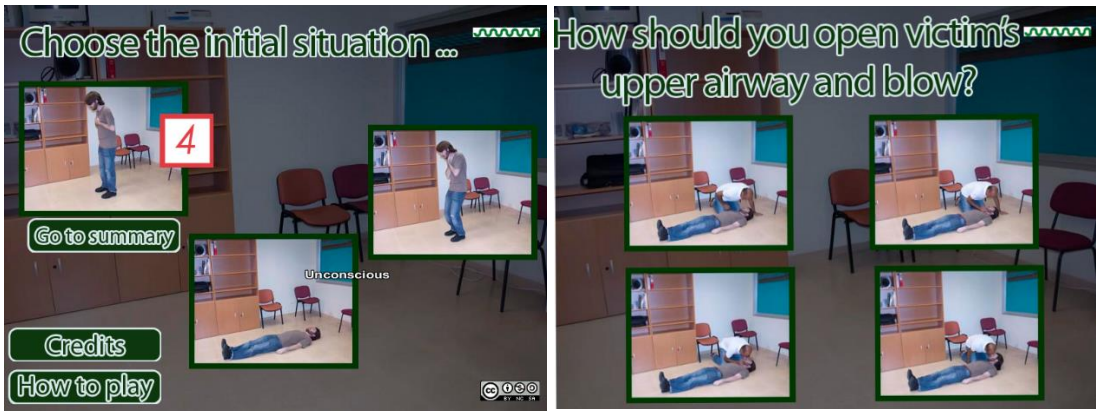

Figure 11. Screenshots of the First Aid Game used in the first case study: the three game levels with a score (left) and visual choices in a level (right).

The game was first developed and evaluated by the e-UCM research group and actual emergency physicians (Marchiori et al., 2012). The game was originally validated with an experiment that included pre-post questionnaires to measure players' knowledge before and after playing, and a control group to compare the game effect against that of a theoretical-practical demonstration by a trained instructor. Players in the experimental group gained, on average, 2.07 points (on a 10-point scale), compared to control group learners who gained 3.61 points. This validation proved that the game achieved its goal of making player learn first aid procedures. The game was later adapted and updated to the Unity 3D videogame engine using *uAdventure* (I. Pérez-Colado, Pérez-Colado, Martínez-Ortiz, Freire, & Fernández-Manjón, 2017), an authoring tool developed by e-UCM. The adaptation also incorporated the collection of game interaction data used for the present case study.

4.2.2. Data captured

The data captured for this case study was gathered in a set of experiments with N=227 high school students from a charter school in Madrid, Spain. We initially conducted a formative evaluation in two sessions with 28 students. With the feedback from these sessions, we tested the remote data collection in the school settings and prepared for the main experience. Out of the remaining participants (N=199), gender was not obtained for 15 students due to an error handling a questionnaire. For the other 184 students, the gender distribution was 46.7% males and 53.3% females. The median age was 14 years old.

Two questionnaires were used in the experiments. The pre-questionnaire elicited demographic variables (players' gender and age); the first aid knowledge questionnaire with 15 multiple-choice questions, also used in the original experiment to validate the game (Marchiori et al., 2012) about the game contents; and a questionnaire with 11 5-

point-Likert questions on game habits obtained from (Manero, Torrente, Freire, & Fernández-Manjón, 2016), and slightly adapted for this experiment. The post-questionnaire consisted of two parts: the same first aid knowledge questionnaire used in the pre-questionnaire (to compare results); and a questionnaire to evaluate the experience, with 5 5-point-Likert questions assessing the experience, and optional free-text sections for feedback. The score for the first aid knowledge questionnaires is defined as the total number of correct answers, therefore, possible scores ranged from 0 to 15 points. Internal consistency of the scale used was ensured when the test was created in the original validation experiment (Marchiori et al., 2012). These questionnaires were a simplification of the ones used in the medical domain and had been previously validated.

From the game, we captured players' interactions including game level starts, game levels endings and scores, selections in every multiple-choice situations and questions, and interactions with game elements (character, phone, defibrillator). All interaction data traces were collected using the xAPI-SG Profile. Figure 12 represents an example xAPI-SG statement collected from the First Aid Game. The statement represents that the player (with anonymous identifier given in the *actor*, *name* field) has selected (*verb*) the correct response 112 (*result*) in the question about the number of emergencies (*object*), at the given timestamp.


```
{
  "actor" : {
     "name" : "XXXX"
  },
  "verb" : {
     "id" : "https://w3id.org/xapi/adb/verbs/selected"
  },
  "object" : {
     "id" : "http://a2:3000/api/proxy/gleaner/games/<game-id>/<version-id>/NumeroEmergencias",
     "definition" : {
        "type" : "http://adlnet.gov/expapi/activities/question",
     }
  },
  "result" : {
      "success" : true,
      "response" : "112"
  },
  "timestamp" : "2017-01-27T03:20:25.571Z"
}
```


Figure 12. Example of an xAPI-SG statement captured from the First Aid Game: the player has selected the correct response (112) in the question about the emergency number.

4.2.3. GLA variables

The xAPI-SG statements were then analyzed to extract higher-level GLA variables by performing aggregations, retaining maximum and first values, or specific responses in some game selections. Table 4 presents the full list of GLA variables derived from the interaction data collected from the First Aid Game.

Table 4. Game Learning Analytics variables derived from interaction data in the first case study (First Aid Game).

Variable Name	Type	Description
gameCompleted	Binary (true, false)	True if learner completed the game; False otherwise
score	Numerical in range [0,10]	Total score obtained in the game
maxScoreCP	Numerical in range [0,10]	Maximum score obtained in "chest pain" level
maxScoreU	Numerical in range [0,10]	Maximum score obtained in "unconsciousness" level
maxScoreCH	Numerical in range [0,10]	Maximum score obtained in "choking" level
firstScoreCP	Numerical in range [0,10]	First score obtained in "chest pain" level
firstScoreU	Numerical in range [0,10]	First score obtained in "unconsciousness" level
firstScoreCH	Numerical in range [0,10]	First score obtained in "choking" level
timesCP	Integer	Number of times student completed "chest pain" level
timesU	Integer	Number of times student completed "unconsciousness" level
timesCH	Integer	Number of times student completed "choking" level
int_patient	Integer	Number of interactions with patient (game character, NPC)
int_phone	Integer	Number of interactions with phone (game element)
int_saed	Integer	Number of interactions with defibrillator (game element)
failedEmergency	Binary (true, false)	True if learner failed, at least once, the question about the emergency number; False otherwise
failedThrusts	Binary (true, false)	True if learner failed, at least once, the question about the number of abdominal thrusts per minute; False otherwise
failedHName	Binary (true, false)	True if learner failed, at least once, the question about the name of Heimlich maneuver; False otherwise
failedHPosition	Binary (true, false)	True if learner failed, at least once, the question about the initial position for Heimlich maneuver; False otherwise
failedHHands	Binary (true, false)	True if learner failed, at least once, the question about the hand position for Heimlich maneuver; False otherwise

The GLA variables provide information about game completion, first and maximum scores per game level, number of tries per game level, number of interactions with the game elements and whether the selections in specific in-game questions were correct or incorrect. These variables were consequently used for the prediction models.

4.2.4. Prediction models and results

The input data for the prediction models included all the variables described in Table 4, for each player. The prediction models, built using RStudio, were additionally created with and without pre-test information as input, to further determine if the pre-test is essential to predict players' knowledge after playing or not. The target variable of the predictions is the post-test score. Two types of models were created: linear models to predict exact score in range [0-15], and classification models to predict pass/fail category (establishing pass as 8 points out of 15).

The prediction models selected included those widely used in the literature for data mining applied to learning analytics data: regression and decision trees, and linear and logistic regression. While trees can show complex, non-linear relationships providing easy-to-understand models, regression is useful when data are not extremely complex or not a lot of data are gathered. Additionally, these models are white-box models, which will allow us to relate the results obtained to our input data to obtain further information related to the traces collected from the game. A priori, our dataset is not too large, so regression should still be viable; however, if complex relationships appear, trees are expected to be better at discovering them. We additionally included two methods commonly mentioned in the literature: Naïve Bayes for classification, and support vector machines for regression (SVR), testing different non-linear kernels (polynomial, radial basis and sigmoid) (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) and tuning the different parameters, with the ranges recommended in the literature (Hsu, Chang, & Lin, 2016). Models were compared using 10-fold cross validation. When predicting pass/fail, and since data were not balanced (169 students passed the post-test, while only 30 failed it), classification models were created with an undersample of 78 students (40% from the fail class, 60% from the pass class) and tested on the original sample. The results of the prediction models are presented in Table 5.

Not all the variables used as input for the models (listed in Table 4) had the same relevance towards predictions. When pre-test information was included, pre-test score appeared among the most relevant variables, but so did the final game score (*score*), and the number of times each situation was repeated (*timesCP, timesU, timesCH*), the maximum score achieved in the "chest pain" game level (*maxScoreCP*) and, the

Table 5. Results of prediction models of first aid knowledge for the first case study (First Aid Game).

Pre-test?	Pass/Fail prediction				Score prediction (scale [0-15])	
	Data mining model	Success measure		Error	Data mining model	Error
		Precision	Recall	MR		Mean (SD)
Yes (*pre+game*)	Decision tree	81.6%	94.2%	16.2%	Regression tree	2.22 **(0.55)**
	Logistic regression	89.8%	**98.3%**	**10.5%**	Linear regression	1.68 (1.44)
	Naïve Bayes Classifier	**92.6%**	89.7%	15.1%	SVR (non-linear kernels)	**1.47** (1.33)
No (*game-only*)	Decision Tree	88.6%	92.4%	17.3%	Regression tree	2.38 **(0.62)**
	Logistic regression	87.2%	**98.8%**	**12.7%**	Linear regression	1.89 (1.54)
	Naïve Bayes Classifier	**89.7%**	90.6%	16.9%	SVR (non-linear kernels)	**1.56** (1.37)

number of interactions with the game character (*int_patient*). Solely with game interactions, the most important variables both predict pass/fail categories and exact score included the number of interactions with the game character, and the first and maximum score achieved in the "chest pain" level. We hypothesized that, as players tend to play the game from left-to-right, the "chest pain" level was commonly played first, therefore, their results on their first level played may have a greater influence on the final knowledge acquisition. Regarding the interactions with the game character, as the game forces players to repeat some situations, a high number of interactions may point to a "trial-and-error" strategy.

4.2.5. Discussion and conclusions

The highly accurate results obtained in this initial case study provide initial evidences that players can be effectively assessed based on their game interaction data. We obtained high accurate predictions of knowledge (as post-test results) from previous information: although the models that included pre-test data found it useful for the predictions, we have verified that similarly high-accurate results were also obtained predicting post-test scores solely from in-game interactions, without pre-test information. Therefore, we can focus in the following to predict players' performance solely from in game interactions.

The encouraging results obtained on this case study suggest that our approach may be generalized at least to other similar cases, such as games for procedural learning or game-like simulations with narrative structure that are quite common in several domains (e.g. military, medicine). Both can provide similar interaction data, and therefore, by following the described steps, a similar approach could be applied.

This case study has some limitations, as the data used are from one serious game and a single school, which could potentially bias the results. However, we consider that

the approach could be generalized for a wider range of games and students with similarly accurate results.

From this first case study, we obtained some initial guidelines for our final evidence-based assessment process. First of all, the interaction data was based on the game design and learning design of the game, containing both game-independent information (game completion) and game-dependent information (number of times per level, based on the design decision to be able to repeat the levels). The definition of most GLA variables was straightforward from the xAPI-SG interaction data (scores, responses, completion) but also adapted to the game characteristics (one variable per each of the three game levels). The accurate results also confirmed our proposal and allowed us to continue with the process, aiming to explore it with different serious games. A final lesson learned is that we were able to relate specific assessment results with the game design only because the final prediction models were interpretable: this allowed us to obtain information about the relevance of the variables and infer players' strategies and relate them to learning outcomes. With black-box models, such a discussion would not have been possible as we would not have had that feedback.

The process, results and details about this case study have been published in (Alonso-Fernández, Martínez-Ortiz, Caballero, Freire, & Fernández-Manjón, 2020), a publication that is included as part of this thesis. For the full text and details of the publication, see subsection 6.1.2.

4.3. Second case study

In this section, we describe the second case study carried out to further test the evidence-based process to assess players and refine the process. To verify the generalization of the approach, we used a serious game with a different purpose (raise awareness about bullying and cyberbullying) and tested a wider set of prediction models.

The following subsections describe the serious game used (4.3.1), the interaction data captured from it (4.3.2), the analysis on that data to create GLA variables (4.3.3), the prediction models used to assess players and the results obtained (4.3.4), and the discussion and conclusions of this case study (4.3.5).

4.3.1. The game: *Conectado*

The game Conectado (Calvo-Morata, Rotaru, et al., 2020) is a serious game to raise awareness about bullying and cyberbullying, developed as part of Antonio Calvo's thesis (Calvo Morata, 2020). Conectado places players in first person as a student that transfers into a new school and, during the first week, becomes increasingly bullied by classmates. The aggressions happen both in the school and at home, where the bullying continues via the mobile phone and the social media (i.e., it becomes cyberbullying). During the 5 in-game days, the player can interact with several in-game characters: parents, schoolmates (which represent different attitudes towards bullying and cyberbullying) and teachers. The game has a linear flow and, depending on the actions taken, such as mentioning the problem to the character's parents or teachers, players will reach one of the three different game endings. By design, player's choices only have an immediate effect on the next dialogues and some of the following actions in the game, but do not affect the main storyline until just before the ending. This ensures that all players will go through all the situations represented in the game, therefore having a similar awareness experience, while still experiencing their actions as meaningful, even while they have minimal effect on the overall flow of the game. Linear play also makes all playthroughs of comparable length and provides all players with a common experience for their in-class post-game discussions. Figure 13 depicts two screenshots of Conectado, showing the player in the school with classmates (left) and the in-game mobile phone, allowing different options to answer (right).

Figure 13. Screenshots of the serious game Conectado, used in the second case study: dialogue with a non-playable character (left) and choices in a conversation in the in-game mobile phone (right).

4.3.2. Data captured

The data used in the case-study was obtained from N = 1109 participants (ages 12-17) from 11 schools around Spain. In all experiments, participants completed a pre-test, a gameplay of Conectado, and a post-test, in that order (Calvo-Morata, Alonso-Fernández, Freire, Martinez-Ortiz, & Fernández-Manjón, 2020). Minimal time elapsed between the gameplay and either of the two tests, and the complete sessions lasted a total of around 50 minutes, fitting in an average-length lecture session in Spain's schools.

The pre-test and the post-test both assess bullying and cyberbullying awareness before and after playing Conectado. The set of questions included in both tests derive from multiple formal and widely accepted questionnaires that have been demonstrated effective in the school population of Spain (Álvarez-García, Núñez Pérez, & Dobarro González, 2013; Garaigordobil & Aliri, 2013; Ortega-Ruiz, Del Rey, & Casas, 2016). In total, the pre-test and the post-test included 18 7-point Likert questions, eliciting how much players agree with each of 18 statements on bullying and cyberbullying. The questionnaire has a Cronbach's alpha of 0.95. The score of each test is calculated as the mean of all answers; therefore, possible test scores range from 1 to 7.

As well as the responses to both questionnaires, the game interaction data traces were collected during the experiments, including in-game days starts and ends, scenes changes, and interactions (characters, objects). All traces were represented using the xAPI-SG Profile. Figure 14 depicts an example xAPI-SG statement collected from Conectado. The statement represents that the player (with anonymous identifier given in the *actor, name* field), has interacted (*verb*) with the game object computer (*object*) at the given timestamp. Additionally, the *result* field contains information about the game day and hour, and that the in-game mobile has messages.

```
{
  "actor" : {
     "name" : "XXXX"
  },
  "verb" : {
     "id" : "http://adlnet.gov/expapi/verbs/interacted"
  },
  "object" : {
     "id" : "http://a2:3000/api/proxy/gleaner/games/<game-id>/<version-id>/Computer",
     "definition" : {
        "type" : "https://w3id.org/xapi/seriousgames/activity-types/game-object",
    },
  "result" : {
      "extensions" : {
        "GameDay" : 1.0,
        "GameHour" : "21:30",
        "MobileMessages" : "True"
    }
  },
  "timestamp" : "2018-05-17T12:04:56.835Z"
}
```


Figure 14. Example of an xAPI-SG statement from Conectado: the player has interacted with the computer in the game. Additional information is encapsulated in the result field.

4.3.3. GLA variables

The xAPI-SG statements were then processed to derive the GLA variables to be used in the analysis. For each type of statement, we stored the following information:

- For "accessed" actions, an identifier for the target, such as "school_bathroom"
- For "initialized" actions, an identifier for the object of the action, such as the full game or a specific in-game day, and a timestamp.
- For "completed" actions, same information to that of "initialized", and, if the full game has been finished, the specific ending reached within the result field.
- For "interacted" actions, the target (which can be an in-game object, when using items, or a character, in the case of conversations).
- For "progressed" actions, an object identifier. For example, when tracking the changes in variables that represent the level of friendship with other characters, the identifiers of those characters are used.
- For "selected" actions, the object and the results of the action to track in-game decisions. For example, when players can choose to mention the ongoing bullying to parents, the results would include the player's choice, and the object would identify the point where that choice was taken.

With this information, we can calculate the values of the GLA variables (Table 6), to be consequently used in the prediction models.

Table 6. GLA variables derived from interactions in the second case study (Conectado).

Variable name	Type	Description
accepted_c, *c* in *[Alison, Guillermo, Jose]*	true/false	Player has accepted a friendship request on in-game computer of character *c*
accessed_bathroom	true/false	Player has accessed the school bathroom
confront_Alejandro	true/false	Player has confronted Alejandro
duration	continuous	Total time playing *Conectado* (in minutes)
duration_day_d, *d* in *[1,2,3,4,5]*	continuous	Total time playing day *d* of *Conectado* (in minutes)
ending_number	categorical	Ending reached by the player: 1 for worst ending, 2 for regular, and 3 for best ending
find_earring	true/false	Player has helped Alison to find her earring
friendship_decrease_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]*	discrete	Number of times the player has decreased the level of friendship with character *c*
friendship_increase_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]*	discrete	Number of times the player has increased the level of friendship with character *c*
gum_washed	true/false	Player has washed the gum from the clothes
has_ended_game	true/false	Player has ended the full *Conectado* game
interactions_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Mother, Father]*	discrete	Number of interactions the player has carried out with character *c*
mock_Maria	true/false	Player has mocked Maria
shared_password	true/false	Player has shared the password with classmates
tattle_to_parents	true/false	Player has mentioned bullying to parents at home
tattle_to_teacher	true/false	Player has mentioned bullying to teacher at the school
used_computer	true/false	Player has used the computer at home
used_friends_app	true/false	Player has used social network app on smartphone
used_mobile_chat	true/false	Player has used instant messaging on the smartphone

4.3.4. Prediction models and results

The prediction models in this case aim to predict the increase in bullying awareness as a result of playing the game. We define the bullying awareness increase as the difference between the post-test mean score and the pre-test mean score for each player. Therefore, this continuous variable is the target variable for prediction models. The GLA variables described in Table 6 were used as input in all the prediction models.

We have used different prediction models to predict the exact value of the increase in bullying awareness and compared predicted results with those obtained in the pre-post tests. As prediction models, we chose linear regression, regression trees, Bayesian regression, Support Vector Machines for Regression (SVR), k-nearest neighbors (k-NN), neural networks, random forests, AdaBoost, and gradient boosting. All models were tested with 10-fold cross validation. For all models, different parameters were tuned to find the best ones. For each of the 9 prediction models, Table 7 shows the mean absolute error (MAE) and the standard deviation (SD) (normalized to scale [0-10]) for the predictions with the best combination of parameters found for that model.

The model that provides the best results is a Bayesian regression, closely followed by a gradient boosting model, with random forests and AdaBoost models at very similar error levels, and all other models providing acceptable results. The difference between the best models is not significant. The variables that were most relevant in the best-

Table 7. Results of prediction models of bullying awareness increase for the second case study (Conectado).

Prediction model	Mean Absolute Error (MAE) *normalized to scale [0-10]*	Standard Deviation (SD) *normalized to scale [0-10]*
Linear regression	0,581	0,047
Regression trees	0,557	0,055
Bayesian ridge regression	0,540	0,053
SVR	0,556	0,051
k-NN	0,578	0,048
Neural Networks	0,557	0,050
Random Forests	0,551	0,052
AdaBoost	0,551	0,057
Gradient boosting	0,548	0,052

performing models included: the number of interactions with the character Jose (*interactions_Jose*), possibly showing that a high number of interactions with any character may be a result of a high immersion of the player in the game; the ending reached (*ending_number*), which is the result of the in-game actions and decisions taken, therefore, it relates to players' behavior in the game, therefore, we consider that an adequate behavior shows higher awareness of players, and this could be related to a higher inclination to be attentive in the game and therefore further increase their awareness; the duration of in-game day 4 (*duration_day_4*), possibly showing that the specific content of that day (threats, theft, and identify theft) may be more relevant to the target group (12-17 years old), impacting their awareness increase; and the duration of in-game day 3 (*duration_day_3*), where above-average durations in these day content (higher presence of social media) may show players losing attention in the game by being distracted by the in-game social media application.

4.3.5. Discussion and conclusions

In this second case study, we have replicated the accurate prediction results with a SG with a different purpose (raise awareness about bullying, and not learning), with a larger dataset and a wider range of prediction models, some more complex ones. From the similarities encountered, we consider that the approach used can be generalized to other serious games or, at least, to other linear, narrative serious games.

From the steps taken, we consider that this process could be generalized to carry out other evidence-based evaluations of the effectiveness of serious games. The steps followed can be generalized, using a standard to track in-game interactions such as the xAPI-SG Profile. Once interaction data are collected, a further step towards generalization is to gather an initial set of variables to derive from the xAPI-SG traces, based on available fields such as the duration of in-game activities, and interactions with relevant in-game items and characters; which can later be complemented with game-dependent information. An initial set of variables can be used as a baseline of what game learning analytics can conclude for the serious game and can be extracted automatically if analytics traces are formatted using the xAPI-SG Profile representation. With those GLA variables, interpretable prediction models can provide information of the relevance of each variable, which can help to interpret and inform the evaluation process and its results. Moreover, using xAPI allow SGs' developers and researchers to build and reuse a tooling ecosystem for both statements gathering, analysis and predictions.

This case study has some limitations to its generalizability. First, the fact that the videogame has a narrative, almost-linear structure and a low playing time restricts the

variability of the interactions for players. Second, the discussion of the relevance of specific variables in our results is limited by the fact that the prediction model is not a black-box model. Finally, the creation and selection of the GLA variables is not straightforward and could limit the generalization of our approach; however, we consider that most of the GLA variables used can be automatically created based on the xAPI-SG fields providing some guidelines and recommendations.

From this second case study, we were able to extract further information for our evidence-based assessment process. This case study enabled us to test our evidence-based approach with a serious game that had a different purpose (raising awareness instead of teaching) and replicating the accurate results obtained, additionally with a wider range of prediction models and a larger sample size, following our pointed-out result that larger number of participants were required in experiments. In the process, we were able to collect most of the interaction data and the GLA variables derived from it simply based on the fields and types available in the xAPI-SG Profile. This showed us the suitability of this profile for different types of serious games. We additionally handcrafted some GLA variables (the specific ending reached) that were relevant based on the learning design of the game. Again, we noticed the suitability to use interpretable prediction models, as the information that they provide about the predictive relevance of the input variables allowed us to relate specific game actions to awareness increase.

The process, results and details about this case study have been published in (Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2020a), a publication that is included as part of this thesis. For the full text and details of the publication, see subsection 6.1.3.

4.4. Evidence-based assessment process of serious game players

In this section, we present the final evidence-based assessment process of serious game players, based on the collection on interaction data to obtain GLA evidences, and predict their learning based on such evidences with data mining techniques. The process is based on the lessons learned from the work carried out and the conclusions extracted in the case studies with the serious game First Aid Game (described in Section 4.2) and the serious game Conectado (described in Section 4.3).

Additionally, this assessment process derives from some previous work on GLA systematization, in which we started to explore how the application of GLA could be standardized (Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2017). This publication is included as part of this thesis. For the full text and details of the publication, see subsection 6.2.1. In that paper, we explored:

1. The use of a standard tracking model to exchange information between the serious game and the analytics platform. This allows reusable tracker components to be developed for each game engine or development platform.
2. The use of standardized analysis and visualization assets to provide general but useful information for any given serious game that sends data in the previously stated standard data format.

In that work we studied the importance of determining the suitable set of GLA variables to obtain rich information from a serious game: a complete set of game-independent variables is recommended to simplify and systematize the process, with the possibility to extend them with some game-dependent variables if needed.

The full evidence-based assessment process of players using serious games is a two-step process described in the following subsections:

- During serious game validation
    - The collection of player data, both pre-post questionnaires and interactions in the game (subsection 4.4.1).
    - The feature extraction process to derive GLA variables from the game interaction data (subsection 4.4.2).
    - The prediction models with GLA variables (subsection 4.4.3).
- During serious game deployment
    - The assessment of players based on interactions (subsection 4.4.4).

The process is carried out in two steps: during the game validation phase, the prediction models to assess players with are created and validated; in the game deployment phase, players' can automatically be assessed based solely on their game interaction data, that is, questionnaires can be completely avoided, simplifying teachers' tasks to obtain a measure of how much effect the game is having on players.

The steps carried out during the game validation phase, explained further in the following subsections, are depicted in Figure 15.

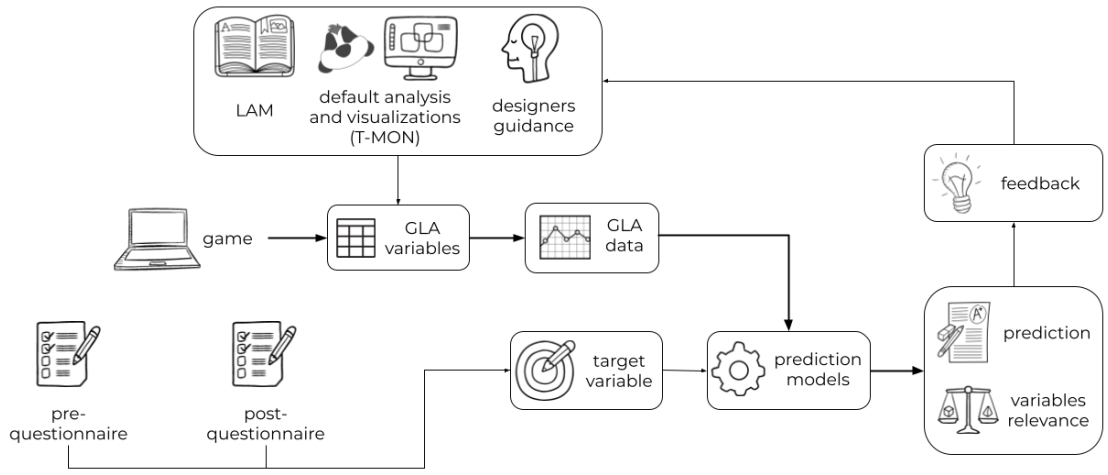

Figure 15. Evidence-based assessment process of players using serious games: the game interaction traces collected fill the pre-defined set of GLA variables to be used as input for the prediction models. The target variable used for prediction is based on pre-post results.

4.4.1. Collection of player data: pre-post questionnaires and game interaction data

The first step in our process is the collection of the data to assess players with. For this purpose, we need to collect both pre-post questionnaires (or any other validated measure to be used as the target value for the predictions) as well as interaction data. Questionnaires should be formally validated by experts in the domain, to ensure that they provide a reliable measure of the characteristic that the game seeks to affect, such as awareness or knowledge.

Although the type of data that can be collected from a serious game will depend on its content, structure and features, there are some common interactions in game analytics (GA) and learning analytics (LA) that can be extrapolated to serious games. GA data information will related to the game design (e.g. number of clicks, avatar location in the game environment and characteristics, movements and changes of scenes or levels, items used, total time spent in the game, interactions with interface elements and non-player characters, points scored, in-game selections, and quest

completions), while LA data will focus on the learning design (achievements, errors made, responses). GLA data comprises both the game and learning design of the games, reflecting information about the learning progress and process of players/learners.

To systematize the specific data to be collected from serious games, we propose the use of the validated and standardized Experience API for Serious Games Profile (xAPI-SG) (Serrano-Laguna, Martínez-Ortiz, et al., 2017), described in detail in section 2.2.1. The use of a standard data format, such as xAPI-SG, is a clear benefit to provide a step further in the systematization of the collection of traces and their analysis to derive relevant information from user gameplays. Standard formats facilitate the integration of tools from different providers and help to comply with personal data-protection laws: art. 20 of the EU GDPR requires data controllers to use a "structured, commonly used and machine-readable format" when users request access to their data, or transfer to other data controllers (European Commission, 2018). Additionally, as we found out in the systematic literature review detailed in section 4.1, standard data collection formats are not commonly reported on the literature, and their uptake would greatly assist in result replication and data sharing. Having a common interchange format also fosters the creation of a tool ecosystem created by different actors.

4.4.2. Feature extraction process: GLA variables from interaction data

Once the raw traces with user interaction data are collected, they can be analyzed to extract higher-level meaningful information about the actions of players within the game, in what is called a *feature engineering* process. Our process synthesizes the information available in the data traces (collected in xAPI-SG format) into a smaller set of GLA variables.

Ideally, the definition of such variables should be described in the game's Learning Analytics Model (LAM) (I. Pérez-Colado, Alonso-Fernández, Freire, Martínez-Ortiz, & Fernández-Manjón, 2018), cooperatively created by both educational experts and game designers. LAMs build on the game's learning design and game design, which define the educational goals of the game and how these are reflected on the specific game design choices taken depending on their educational goals. Based on both designs, a LAM determines the data to be collected from the game and how these data are to be analyzed into GLA variables and interpreted to provide meaningful information about the actions of a player in the game. It also may define any posterior visualization, feedback or reporting to do with the analysis results.

If such a LAM is not available, the definition of the GLA variables can be based on:

- Game designers' suggestions about what information to obtain from the game and analyze it into GLA variables.
- Using the xAPI-SG as the data collection standard, we can provide a default set of GLA variables to be easily extracted from the fields available in the Profile. While game-specific variables as specified in a LAM or suggested by expert designer knowledge are of course preferable, a set of ready-to-use generic variables can be highly useful to complement game-specific variables, and allows the use of our process even when no LAM or designers are available. Table 8 details the xAPI-SG fields and the GLA variables that can be obtained from the fields, providing a non-exhaustive set of pre-defined GLA variables for each player that can be easily derived from any set of traces that follow the xAPI-SG Profile. Such variables include the number of interactions with each in-game object and character (count of interacted traces per object), or the duration of each level/game (difference in timestamp of completed and initialized traces per object of type serious-game or level).
- Analysis and visualizations of the xAPI-SG traces can provide important insights on the data collected and guide the choice of some GLA variables. For the latest purpose, we have created our data science environment called T-MON (a trace monitor in xAPI-SG format), detail below.

T-MON: Monitor of traces in xAPI-SG

To support the feature extraction process, we have developed T-MON (Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2021), a monitor of traces in the standard xAPI-SG. T-MON contains a set of Python Jupyter Notebooks that provide a default set of analyses and visualizations that can be applied to any given JSON file containing xAPI-SG traces: overall game progress; choices in alternatives, and if applicable, those considered correct and incorrect; progress, scores and times per game activity or subsection; content seen and skipped; and interactions with game items and areas and over time. The interactive interface allows to filter the data and configure the visualizations to gain a more in-depth insight into the data (Figure 18, right). T-MON is intended both to provide quick overviews of collected data and to allow in-depth exploratory analysis to refine the choice of GLA variables that will be used in subsequent steps: the Jupyter Notebooks (Project Jupyter, 2020)

Table 8. Correspondence of xAPI-SG traces (object type, verb and other fields) to derive GLA variables.

xAPI-SG fields			GLA variables	
Object type	**Verb**	**Other fields**	**Name**	**Description**
Accessible: area, cutscene, screen, zone	Accessed	Object id	*Accessed_id*	Number of times the accessible *id* has been accessed
	Skipped	Object id (cutscene)	*Skipped_id*	Number of times the cutscene *id* has been skipped
Completable: serious-game, level, quest	Initialized	Object id, timestamp	*Duration_id*	Duration of completable *id* (calculated in combination with *completed* trace of same *id*)
	Progressed	Object id, result progress, timestamp	*Progress_id_time*	Progress in completable *id* per timestamp *time*
	Completed	Object id	*Completed_id*	True if completable *id* has been completed
		Object id, timestamp	*Duration_id*	Duration of completable *id* (calculated in combination with *initialized* trace of same *id*)
		Object id, result score	*Score_id*	Score obtained in completable *id*
Alternative: question, dialog-tree, menu	Selected	Object id (question), result success	*Correct_id*	True if question *id* has been successfully answered
		Object id (dialog), result response	*Response_id*	Response selected in dialog *id*
		Object id (menu), result response	*Selection_id*	Option selected in menu *id*
Target: non-player character, enemy, item	Interacted	Object id	*Interactions_id*	Number of interactions with target *id*
	Used	Object id (item)	*Uses_id*	Number of uses of item *id*

T-MON builds upon are a commonly used tool in data science to perform such analyses and provide access to an extensive and actively maintained collection of utilities to manipulate and explore data (Jupyter Team, 2020). The variables included in Table 8 also constitute a good starting point for refinement using T-MON.

Some of the visualizations included in T-MON are depicted in Figure 16 (left to right, top to bottom): pie chart with percentage of serious games started and completed; line chart with progress (y-axis) of each player in the game over time (x-axis); bar chart

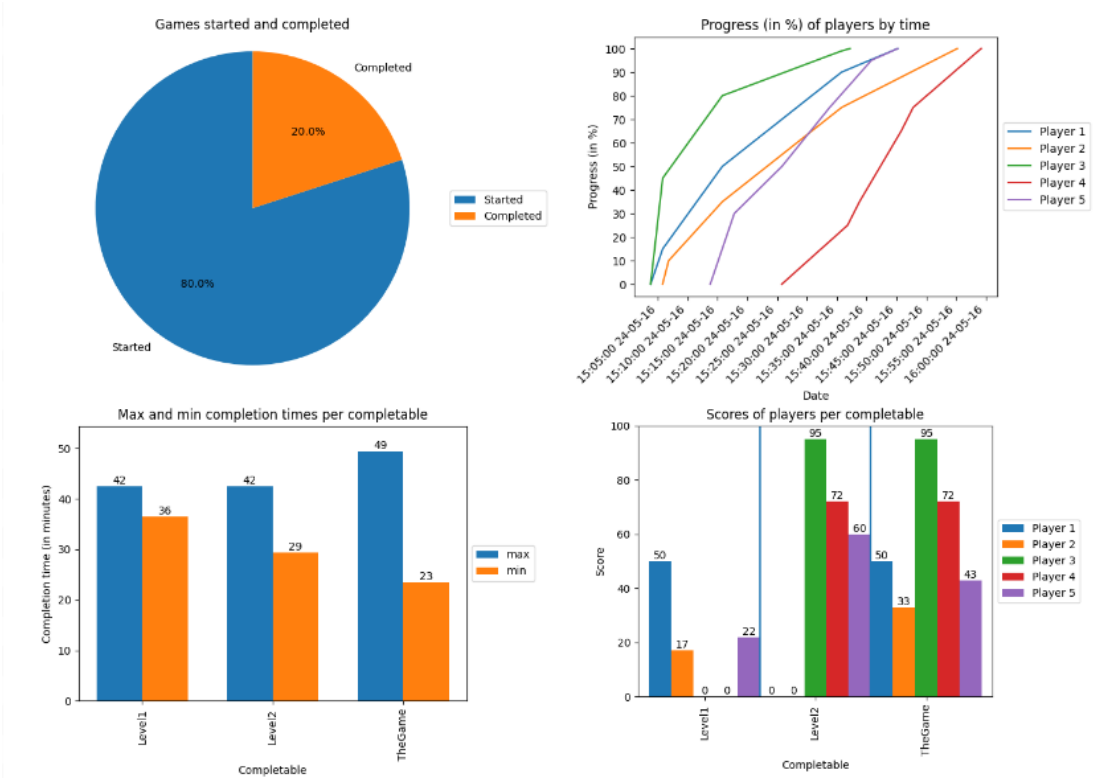

Figure 16. Four of the default visualizations included in T-MON presenting information about games completion, progress, completion times and scores in completables.

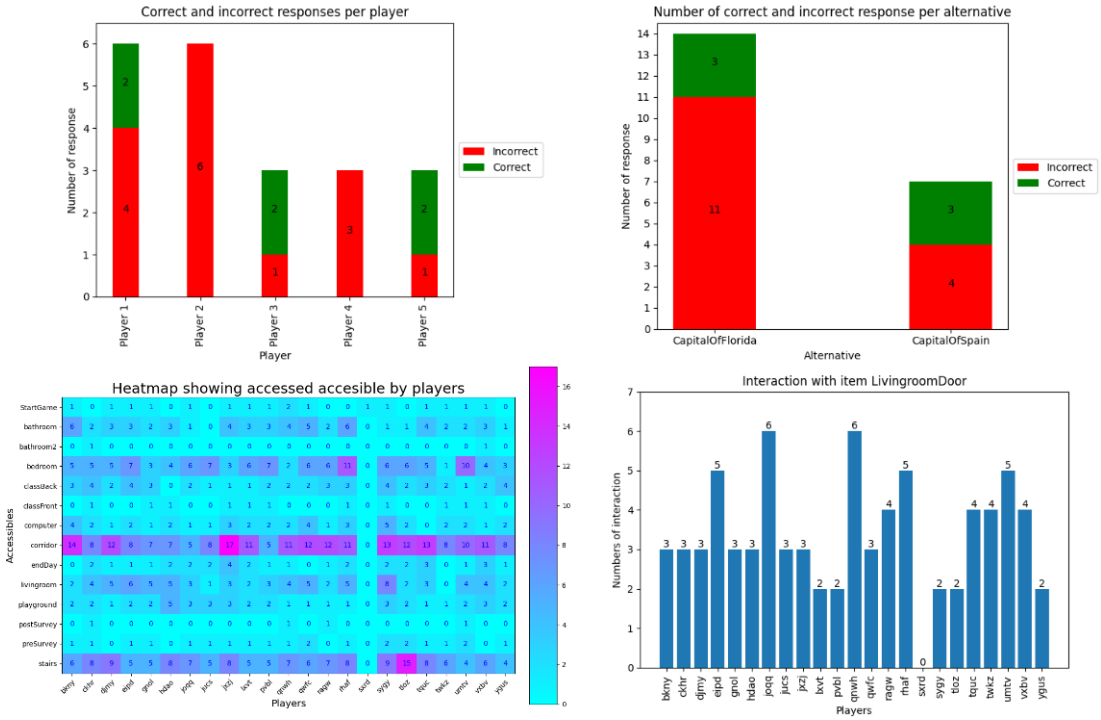

Figure 17. Four of the default visualizations included in T-MON presenting information about correct and incorrect responses in alternatives per player and per question, accessibles and interactions.

with maximum and minimum completion times (y-axis) in each completable (x-axis), max and min times corresponding to each bar per completable; and bar chart with scores (y-axis) obtained by each player in each completable (x-axis), each bar per completable corresponding to one player; and in Figure 17 (left to right, top to bottom): bar chart with correct (in green) and incorrect (in red) number of responses (y-axis) in alternatives per player (x-axis); bar chart with correct (in green) and incorrect (in red) number of responses (y-axis) per alternative (x-axis); heatmap with times each accessible (y-axis) has been accessed per player (x-axis); and bar chart with number of interactions (y-axis) per player (x-axis) with an item.

T-MON is open-source and freely available in a GitHub repository[2] providing information about the use of the tool and the analysis and visualizations provided (Figure 18, left).

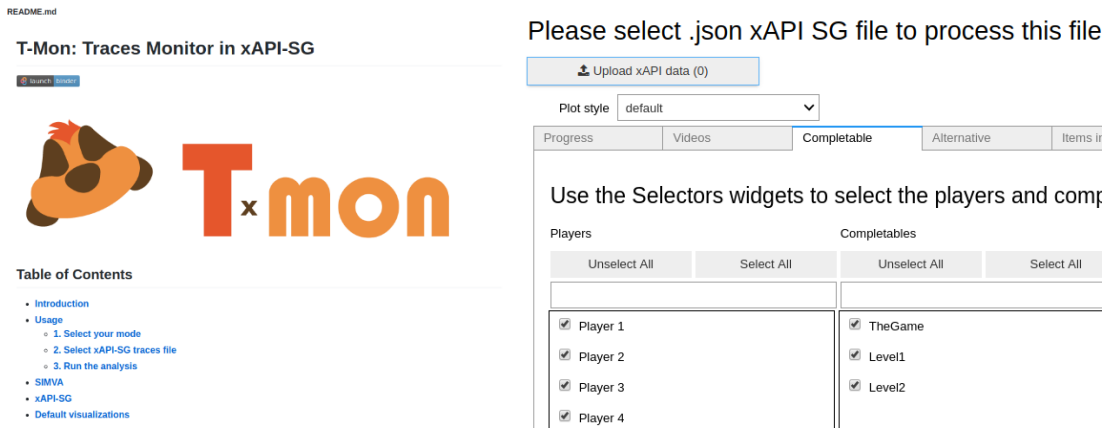

Figure 18. T-MON main GitHub repository page (left), and interface with configuration options (right).

The publication describing T-MON is included as part of this thesis. For the full text and details of the publication, see subsection 6.2.2.

Once the GLA variables have been chosen, synthetizing the information obtained from users' interactions, they can be used to predict the serious game's effect on players, as described in the following section.

4.4.3. Assessment prediction with GLA evidences

The next step is to create the prediction models to accurately measure the effect of the game on its players. To consider a serious game effective in educational scenarios, it first needs to be validated, ideally using a formal validation process (like pre-post

[2] https://github.com/e-ucm/t-mon

questionnaires). We use the formal validation step to create the prediction models that will be used in the deployment phase. During these experiments, we also collect relevant GLA data from players' in-game interactions. The prediction models then have:

- As input, the GLA variables filled with the data collected from players' interactions with the game.
- A target variable, the effect cause by the game on players. By default, that is the improvement in score (difference between pre- and post- questionnaire results, as in the second case study), but if we were only interested in measuring the final effect on players after playing, the post questionnaire score alone could be used as the target variable (as in the first case study).

This process is experimental and can be iterated until accurate-enough models are created, by changing and refining the GLA variables according to their relevance as reported by the results of the prediction models. In our experiments, we have found accuracies above 90% to be achievable, and suggest this figure as a workable goal. Once an accurate-enough level is reached, the final prediction model is retained for the next step of deployment, where it will be used for automatic non-intrusive assessment of players based solely on their interaction data.

For the specific prediction models to be tested, an increasingly broad and varied range of options is available. At least in the first iterations, we recommend using interpretable models that provide information about the relevance of the input variables towards the predictions. These xAI models (Adadi & Berrada, 2018) will provide feedback about the importance of specific GLA variables (and, therefore, about users' interactions) for the predictions, allowing to improve the process before moving to deployment. As exemplified in the two case studies, the use of such interpretable models also allows to relate the assessment results to specific behaviors or choices made by players, better understanding their learning process and providing feedback.

Linear and tree-based prediction models are a simple baseline to start from. More complex models may improve the results: for instance, ensemble methods based on trees, such as random forests or gradient boosting. These complex models could provide more precise results while still giving feedback about how relevant the input variables are towards the prediction results. The models may then be reused and adapted for different contexts. Traces can be re-examined to generate additional GLA variables or change existing ones based on variable relevance as reported by such models.

As for the number of users to include in this validation phase, considering the reported number of users in other data-based research on serious games (as described in our systematic literature review, section 4.1), we recommend including at least 100 users. The information gathered during the validation phase can also be used to improve the game or adapt it to players' characteristics for a better learning experience.

4.4.4. From game validation to game deployment

As we have described, our full evidence-based assessment process involves the steps of the serious game validation (collecting both questionnaires and GLA data), and the serious game deployment where assessment is simplified and solely based on players' interaction data.

Once the serious game has been formally validated, the deployment phase can start, with the game applied in classrooms and other real-world educational settings. To be able to gather information from users' experience and to assess them based on their interactions, this application should include the collection of data from relevant interactions. The deployment process for large-scale scenarios reads as follows:

1. **Gameplay:** Students access the SG and play the game from beginning to end. We have used anonymous identifiers that allow only teachers to de-anonymize student data to ensure that privacy requirements are met while still linking questionnaire responses to each student's game-interaction data.
2. **Data collection**: A tracking component integrated in the game sends the relevant traces generated from player interactions to the analytics tool while students are playing. The user interaction traces should follow a well-defined format (e.g. xAPI-SG), as required by the analytics tool that will receive it.
3. **Feature extraction**: The analytics tool takes interaction data as input, and uses it to fill the values of the pre-defined GLA variables. These variables are then used as input to the previously created prediction models, to derive prediction outputs for the students' assessment.
4. **Assessment**: Once students have finished playing, teachers will receive the predicted score based on each students' in-game interactions. They can then use this information, together with any other evaluation of their own, to obtain the final students' assessment.

Note that the prediction models provide the assessment output for students once they have finished playing the game, and therefore once all the input data required by the models is available.

The assessment obtained with this process is therefore automatic and non-intrusive, and simplified from both ends: teacher preparation and execution times typically required for post-game assessment are removed, and students will simply play a game without the added time, disruption, and pressure of completing the questionnaires. Game-based assessment can also provide institutions and managers means of evaluating the efficacy of games for education and simplify the assessment process of their students. The previously used pre-post questionnaires are no longer required during the large-scale deployment, which simplifies the application of SGs in real-world educational settings. This allows students to play the game for longer periods, and/or teachers to include additional activities related with the gameplay (e.g. discussion, post-game questions), instead of the traditional student assessment.

Figure 19 depicts the deployment phase of a serious game using our evidence-based assessment process, once questionnaires are no longer required (note the differences with the game validation phase, depicted in Figure 15).

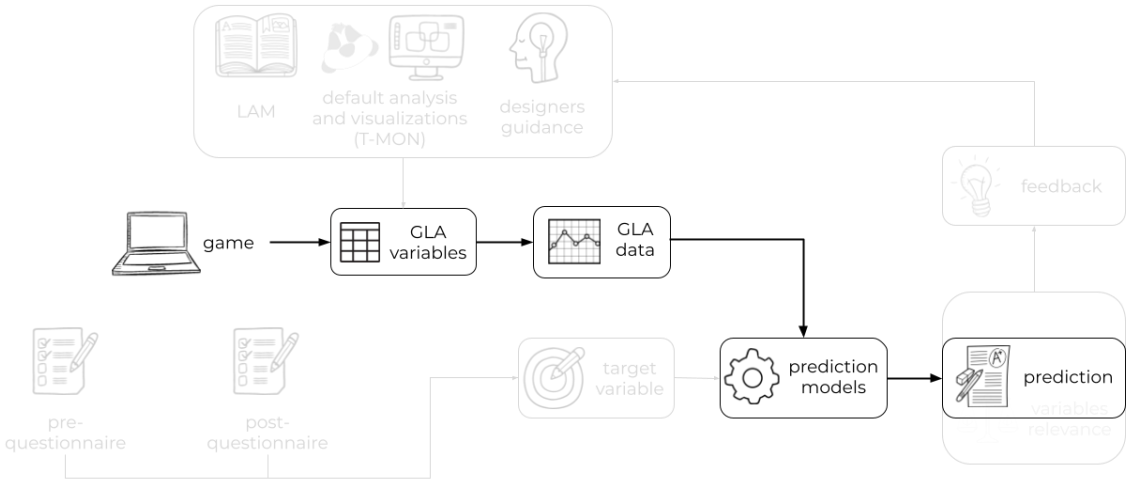

Figure 19. Evidence-based assessment process of serious game players: after validating the game and the prediction models, during the game deployment, players are assessed solely based on their game interactions.

The process, results and details about this full evidence-based assessment process have been published in (Alonso-Fernández, Freire, Martínez-Ortiz, & Fernández-Manjón, 2021), a publication that is included as part of this thesis. For the full text and details of the publication, see subsection 6.1.4.

4.5. Discussion

The evidence-based assessment process presented, and the case studies that exemplify it, have some limitations. The full process is closely tied to the use of the xAPI-SG Profile. This standard may fail to capture the particularities of specific serious games – however, the standard was created after a study of the common interactions in serious games, so we consider that it is a suitable baseline to define general information that can be extracted from a majority of serious games. The games used in the case studies do not cover the variety and complexity of the serious games that exist. To try to tackle these issues, we have selected two serious games that, at least, have different goals (teaching knowledge and raising awareness), with different mechanics. Still, both serious games are similar in their narrative structure and the importance of the dialogues and selection of choices made by players in alternatives. Therefore, we consider that our approach could yield similar results at least in similar narrative serious games. Further research is needed to explore this approach with different types of serious games. As of the participants includes in the experiments, we have tried to ensure larger samples sized to tackle the identified gap in the literature. However, there are also some limitations regarding their characteristics: the participants in the first case study were from the same school, which could bias the results. In the second case study, participants were from different schools in the same country – Spain.

Regarding the steps of the evidence-based assessment process, we have highlighted what we consider to be the key step: the selection of GLA variables from game interaction data (feature extraction process). We consider that this step is essential as we consider, like other authors have pointed out, that the selection of variables is of higher importance for the results than that of the prediction models. To try to support this key step, we have provided both a default set of GLA variables (derived automatically from using the xAPI-SG standard), as well as the exploratory analysis tool T-MON, to obtain a richer insight into the collected interaction data traces, to derive further game-dependent GLA variables if needed. For this step to yield the best results, it is of course preferable that analytics have been included from the early stages of the serious games design. This simplifies all the steps as, as well as analytics are been design, they could also imply changes in the game design to obtain the desired information. Finally, for the prediction models included, we have recommended the use of explainable models: black box models that provide no information could improve the accuracy of the assessment but in earlier stages of the process it could be preferable to use xAI models with information about the predictive relevance of the included GLA variables to obtain feedback and improve the process.

The work carried out in the thesis is based on some previous exploratory work about the applications of GLA for assessment with serious games. In particular, we had explore how LA for SGs can provide insight to improve the application of serious games in different stages of serious games lifecycle (Alonso-Fernández, Cano, et al., 2019; Alonso-Fernández et al., 2018) including validation of the design of a serious game, improvement of the deployment of serious games, and assessment of players using serious games. In that work, that combined three different scenarios of application of LA in SGs, we obtained some lessons learnt that included:

- The importance of using a standard format to collect the data, to simplify integration with other systems (e.g. real-time information analytics system), compare information from different games and even reuse and share the collected data for research purposes.
- The wide range of purposes to apply GLA data with serious games: validation of game design, simplify deployment, and assessment, and provide further information about students' actions in game, engagement and motivation.
- The different stakeholders that can benefit from such applications: game designers and developers to simplify validation of their designs; teachers and educators, to simplify application of games in classrooms; and students/learners to be effectively assessed based on their actions.

The process, results and details about these initial exploratory works were initially published in (Alonso-Fernández et al., 2018), a publication that is included as part of this thesis (for the full text and details of the publication, see subsection 6.2.5), and then extended in the publication (Alonso-Fernández, Cano, et al., 2019), a publication that is included as part of this thesis. For the full text and details of the publication, see subsection 6.1.5.

The application of GLA for serious games does not only benefit the assessment of players: the information obtained with GLA data can improve the serious games, in all steps of their life cycle (from design and development, to validation and deployment). We also explored these opportunities in some initial works of this thesis. The integration of GLA with a game authoring tool, including the collection of GLA data, and its analysis and visualization using a real-time information analytics system can simplify the deployment and application of serious games in educational scenarios (Alonso-Fernández, Rotaru, Freire, Martínez-Ortiz, & Fernández-Manjón, 2017). For that, we proposed the integration of GLA in the game authoring tool uAdventure, the use of the standard xAPI-SG to standardize the data collection, and a default set of analysis and visualizations for the main stakeholders involved. Additionally, this

possibilities can help to improve serious games lifecycle at two stages: while games are in play, providing real-time information to teachers and students in the form of dashboards or warning messages, the deployment of games can be more adapted to each player's needs and progress; while, after gameplays are finished, further analysis can provided deeper information from the gathered data (Alonso-Fernández, Pérez-Colado, et al., 2019). The latest includes, of course, the application of GLA data for assessment purposes.

The process, results and details about these works are published in (Alonso-Fernández, Rotaru, et al., 2017), a publication that is included as part of this thesis, and whose full text and details are included in subsection 6.2.3, and in (Alonso-Fernández, Pérez-Colado, et al., 2019), a publication that is included as part of this thesis, and whose full text and details are included in subsection 6.2.4.

The evidence-based assessment process includes the lessons learnt and conclusions obtained from those initial exploratory works. The described process aims to provide some guidelines to carry out assessment of players using serious games, based on standards (the data collection format xAPI-SG) and with tools to support the essential process of creation and selection of GLA variables (with the exploratory tool T-MON). With these guidelines, we hope to simplify the assessment process of players with serious games, which can be adapted and personalized for each serious game and scenario. Still, we consider that the described process can help to improve the assessment method of players with GLA data and data mining techniques. With this simplification, the application and deployment of serious games in educational scenarios could be fostered with clearer evidences on their impact on players.

Chapter 5. Conclusions, contributions and future work

This chapter presents the final conclusions of the thesis, a summary of all the contributions, and the identified limitations of this work. This chapter also introduces some of the possible lines for future research in related fields to the work of the thesis.

5.1. Conclusions

The assessment of serious games' players has been traditionally conducted with external questionnaires that fail to assess players based on their actual actions while playing the games. However, to increase the application of serious games in educational scenarios, traditionally limited to an additional motivational activity with no impact in the evaluation of students, we consider it essential to improve the use of serious games to assess players, providing accurate measures of how they affect their players. Game Learning Analytics application provide a richer and deeper information about players' actions, progress, and results within serious games, allowing to detect players' strategies and behaviors, creating players profiles and adapting the game experience to each player. The potentially large-scale and rich information gathered from game learning analytics data can be further analyzed with complex data mining techniques. This combination offers new opportunities to improve serious games' players assessment. From the literature review of data science applications to game learning analytics data in serious games, we identified some challenges pointed out by authors: for instance, the few studies reporting evidences to effectively assess students, the lack of standardization in the assessment process, the lack of standards in the collection of interaction data and the limited number of participants in many studies.

Based on that background, we proposed an evidence-based assessment approach that combines the collection of game interaction data, analyzed into GLA data variables, with data mining techniques to obtain accurate predictions of serious games' impact on players. The approach proposed aims to be a step forward to simplify players' assessment using serious games. The use of standards (in our case, xAPI-SG) is a clear benefit to simplify the collection of game interaction data, its analysis into GLA variables, and the sharing of data and replicability of the results. Through using the standard data format, it is possible to define a set of game independent GLA variables,

providing a baseline of the information that can be extracted from many serious games by default (as the xAPI-SG Profile was also created based on the most common interactions present in serious games). To further support the key step of feature extraction, we developed the exploratory tool T-MON to help in the definition and selection of further GLA variables, based on the collected interaction data in xAPI-SG format.

The complete evidence-based assessment process has been exemplified in two case studies with SGs that have different purposes (teaching first aid procedures, and raising awareness about bullying and cyberbullying) and with different prediction models. In each case study, we have showcased how the game design and learning goals were turned into game mechanics and interactions, the game interaction data collected and how it was encapsulated in the xAPI-SG statements, and how the default game-independent GLA variables, together with some defined game-dependent ones, were derived. In both cases, the different prediction results have been highly accurate in predicting the effect of each game on players, and we have been able to obtain measures about the relevance of the variables towards the predictions. This allowed us to discuss the results, and provide some plausible explanations of how each variable affected the assessment, based on players' strategies and actions. The case studies also show how the general evidence-based assessment process can be adapted for particular serious games, while keeping most of the steps as general as possible to increase replicability. We consider that, although each serious game will have its particularities, the process is general enough so that it can provide a reliable baseline (also fostered by the application of the xAPI-SG standard) to use with different serious games, or at least with similar narrative-based serious games.

The approach has some limitations. The evidence-based assessment process aims to be general to increase replicability but this, of course, hinders the efficacy in each particular case as important game-dependent information may be omitted. For that purpose, we have further designed and developed T-MON, a tool to support the key step of selecting GLA variables, by showing particularities of the data that may have been overlooked by the default game-independent variables. The T-MON tool and the default GLA variables are based on the xAPI-SG Profile. If some relevant game interactions cannot be effectively collected with the types and fields available in the Profile, again they could be omitted. This should be fixed during the definition of the game interactions to be collected from the game: the extension field available in the Profile could, in many cases, suffice to collect other relevant information from the game. For more complex situations, it could be necessary to define game-specific verbs

or activity types to collect the data. However, we consider that the xAPI-SG Profile is a good baseline, as it was defined based on the creation of a general interaction model stating the most common interactions in serious games. The serious games used in the case studies are also limited in their goals and characteristics, but being different enough, we consider that they provide some guidelines on how to apply the evidence-based assessment process in different scenarios.

The proposed approach further synthetizes the lessons learned in our previous work in two H2020 European projects, RAGE and BEACONING, regarding the large-scale deployment of serious games including the collection and analysis of game learning analytics data. That perspective has helped us to focus on the systematization of players' assessment, using standards, and simplifying the tasks for teachers and educators to obtain evidences about how much players are learning with serious games. The work carried out during the research stay at Florida State University has also allowed us to contrast our approach with an expert research group in the field of *stealth assessment*, emphasizing the need to systematize players' assessment, and the opportunities that the collected game learning analytics data provide to validate different parts of the game design.

With the steps carried out to systematize the assessment of serious games' players, we can contribute to ease the application of serious games in different educational contexts, providing all stakeholders involved with more accurate and data-based evidences on games effect on players.

5.2. Contributions

The main contributions of this thesis are centered around the systematization of the assessment of serious games' players, using game learning analytics data, standards, and data mining techniques. With the systematization of this evidence-based assessment process, we aim to simplify the deployment and application of serious games in a wider range of educational scenarios, by providing accurate and data-based information that can better prove their efficacy. With this information, teachers, educators and institutions will have further evidences of the effect of games on players, and this can contribute to students/players being more effectively assessed while using these learning tools. The application of serious games in educational scenarios could be fostered following systematic approaches like the one presented, and increase their current limited role as simple motivational activities with no impact on students' final evaluations.

The main specific contributions of the thesis are:

- A **systematic literature review** about data science applications of GLA in serious games: this review provides a detailed overview of the different purposes for which studies have applied data science to the game learning analytics data collected from serious games, the different data analysis techniques applied to such GLA data, the stakeholders targeted by such applications and the varied results obtained in the studies. The review additionally provides information about the serious games used and their purposes, the target participants and the sample sizes included in the studies, and the interaction data collected from such gameplays, and their format. After presenting the results of the studies reviewed, we further point out areas for improvement and recommendations for future research on the area, such as the need to increase the average sample size of participants and use standards for the interaction data collected to simplify replicability of results and data sharing.
- A **full assessment process** of players using serious games based on the collection of interaction data and its analysis with data mining techniques: the described process provides a step-by-step methodology to effectively assess players using serious games based solely on their game interaction data. The evidence-based assessment process details all the steps from game and learning design, the interaction data to be collected, and the standard data format xAPI-SG recommended to collect it, the feature extraction process into GLA variables, and the prediction models to simplify assessment. The process comprises two steps: (1) the initial game validation phase is used to create and validate the prediction models, collecting both traditional formally-validated pre-post questionnaires and game interaction data; and (2) in the final game deployment phase, after the game and the predictions models are validated, assessment is simplified as players can be automatically assessed solely from their interaction data. This way, teachers are provided with simplified tools to assess their students, without relying on external questionnaires. The process also includes the use of standards: the xAPI-SG Profile to collect the interaction data is recommended as it simplifies both the definition of the interaction data to be collected, as well as the feature extraction process into GLA variables, a large number of which can be directly defined based on the fields and types available in the standard.
- **T-MON, a monitor of xAPI-SG traces,** to support the evidence-based assessment process: the T-MON exploratory tool helps in the essential step of the feature extraction and selection of GLA variables from the game interaction data, as it provides an exploratory interface with ready-to-use analysis and

visualizations of the interaction data collected in the xAPI-SG format. The default set of analysis and visualizations included in T-MON provide a deeper insight into the data collected, to define additional GLA variables, beyond the ones created by default following the fields in the standard. Additionally, T-MON simplifies the work of data scientists with standard-based GLA, as it decreases the cost to learn these techniques, allowing them to work in a familiar data science environment, without needing to be experts in xAPI or GLA.

- **Two case studies** that exemplify the evidence-based assessment process: the case studies are based on two serious games with different goals (learning knowledge and raising awareness), and using different predictions models (from classic and simple prediction models to more complex models). The case studies describe in detail all the steps carried out from the game design and learning goals, the game interaction data collected using the xAPI-SG standard format, the creation of both by-default and game-specific GLA variables, the predictions models used, and the high-accurate results obtained. The discussion of the results also served to relate predictive relevance to the game design, uncovering how some players' behaviors relate to learning.
- Additional **publications** that have resulted from the work conducted in this thesis: 5 journal publications (Alonso-Fernández, Cano, et al., 2019; Alonso-Fernández et al., 2020a; Alonso-Fernández, Calvo-Morata, et al., 2019; Alonso-Fernández, Freire, et al., 2021; Alonso-Fernández et al., 2020) and 5 conference publications (Alonso-Fernández, Calvo-Morata, et al., 2017; Alonso-Fernández, Pérez-Colado, et al., 2019; Alonso-Fernández, Rotaru, et al., 2017; Alonso-Fernández et al., 2018; Alonso-Fernández, Calvo-Morata, et al., 2021); as well as other related publications published during these years that do not belong to the core of the thesis (Alonso-Fernandez et al., 2020; Alonso-Fernández, Perez-Colado, et al., 2019; Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2020b).

5.3. Future work

To continue verifying the suitability of our evidence-based approach to assess players of serious games, a clear line of future research is to test the assessment approach in different and more varied contexts.

First, the assessment process could be applied with different serious games: both with serious games that have different educational goals than the ones tested (e.g. changing players' attitudes or behaviors towards some issues) or serious games that have the same goals but in different contexts (e.g. games that aim to teach different knowledge,

or raise awareness about other social issues). This application includes the integration of the assessment process with other already-developed serious games, to use our evidence-based approach to assess their players. Regarding the contexts of application for other serious games, one of the fields that we have started to explore, and we aim to continue to do so, is the application of serious games to educate about gender equality. In this area, we have started to study the applications of serious games to raise awareness about different gender inequalities, sexist behaviors, etc. As many of these inequalities and behaviors start during childhood, when games are commonly used by children and teenagers, serious games seem like a particularly suitable tool to raise awareness about these topics.

Additionally, the assessment process could include testing different prediction models: in the case studies, we have included some classical and simple prediction models, as well as some other more complex models. The field of data mining is in constant change, therefore new prediction models are currently being created and tested. For instance, some studies are exploring new prediction models that only include few training data points (Sucholutsky & Schonlau, 2020; Wang, Zhu, Torralba, & Efros, 2018). New predictions models could be tested to improve the assessment results.

A further step will be to embed the assessment process in the development of a new serious game: considering the assessment process while creating a new serious game, to fully integrate all the steps of the assessment process from scratch during the design and development of the game. This way, all the suitable game interaction data to be collected, analyzed and used during the assessment will be clearly defined from the beginning of the design process (this aligns with the recommendation made by several authors, as described in the literature review, that the interaction data for assessment should be defined early on during the game design process). Moreover, these steps could be iterated as needed while the game is still being developed, extending the interaction data collected from the game or changing the game design to provide other educationally relevant data.

The assessment process could also be integrated with some of the current set of tools available in our research group: uAdventure and SIMVA. uAdventure (I. Pérez-Colado et al., 2017) is an authoring tool to create serious games, including geolocalized capabilities and learning analytics. The steps of the assessment process presented could help game designers using uAdventure to improve the analytics metrics collected, defining their game-specific ones, besides the default set of analytics provided by uAdventure.

SIMVA (I. J. Pérez-Colado et al., 2019) is a tool to simplify the experiments with serious games, managing students groups, anonymization, questionnaires and interaction data. With the collected xAPI-SG interaction data in SIMVA, we could provide the prediction models to assess students: integrating SIMVA with a data science tool such as Python Jupyter notebooks, with the previously-created prediction models during the game validation phase, the interaction data collected from each player could be analyzed, the GLA variables derived, and with the predefined prediction models, obtain an assessment measure for players after they have completed their gameplays. This will automatically simplify the assessment of players in educational scenarios.

Following some of the steps of our assessment approach, and those of the field of *stealth assessment*, the techniques applied can be used to further explore the relationship between players' actions in specific contents of the game (e.g. learning supports and incentive systems) and learning outcomes and performance. In this regard, during the research stay carried out as part of this thesis in Florida State University from February to May 2020, we conducted two studies that have been published in two additional JCR journal publications (Rahimi et al., 2021; Yang et al., 2021).

The xAPI standard is currently being prepared and updated for IEEE standardization. An additional future line of work will be to adapt the process to this new version of xAPI as an IEEE standard once it is completed. Moreover, an extension of the xAPI-SG Profile for the particular case of geolocalized serious games was also proposed; the steps in the evidence-based assessment process could be adapted to include the types and fields of this version of the Profile to improve the assessment when using geolocalized games: these games are particularly adequate in the current pandemic situation, as they allow learning experiences to be conducted outdoors and keeping social distance.

The exploratory tool T-MON can also be extended with new functionalities: so far, the tool is highly tied to the xAPI-SG Profile, providing default analysis and visualizations that are restricted to the fields and types available in the standard. However, the tool could be further extended to include other game-specific information, with an interface that allows users to specify other fields and types that are included in their data traces, and how to analyze them (e.g. providing a default set of analysis and visualizations to select from). The current interface could also be improved by extending the current set of visualizations (e.g. statistical visualizations like boxplots) or providing further configuration options.

Chapter 6. Publications

6.1. Journal publications

This section contains the journal publications included in this thesis. The following subsections present in detail each publication, full citation and impact metrics, abstract and full text of the publication. As an overview, the journal publications included in the thesis are the following:

1. **Applications of data science to game learning analytics data: a systematic literature review:** this publication presents the systematic literature review carried out about the applications of data science techniques to game learning analytics data from serious games. The process and results of this work are included in this thesis as part of the related work section and in the results, in subsection 4.1.
2. **Predicting students' knowledge after playing a serious game based on learning analytics data: A case study:** this publication presents the first case study carried out to test the evidence-based assessment process with the serious game First Aid Game. The process and results of this work are included in this thesis as part of the results, in subsection 4.2.
3. **Evidence-based evaluation of a serious game to increase bullying awareness:** this publication presents the second case study carried out to test the evidence-based assessment process with the serious game Conectado. The process and results of this work are included in this thesis as part of the results, in subsection 4.3.
4. **Improving evidence-based assessment of players using serious games:** this publication presents the final evidence-based assessment process obtained after the two case studies, to assess players based on their interaction data with serious games, using game learning analytics data and data mining techniques. The process and results of this work are included in this thesis as part of the results, in subsection 4.4.
5. **Lessons learned applying learning analytics to assess serious games:** this publication presents an overview of lessons learned applying learning analytics to assess serious games in different contexts and with different purposes. The process and results of this work are included in this thesis as part of the results, in subsection 4.5.

6.1.1. Applications of data science to game learning analytics data: a systematic literature review

Full citation

Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2019): **Applications of data science to game learning analytics data: a systematic literature review.** Computers & Education, Volume 141, November 2019, 103612. DOI: 10.1016/j.compedu.2019.103612.

Impact metrics: JCR 2019, Impact Factor: 5.296, Q1 in Computer Science, Interdisciplinary Applications.

Abstract

Data science techniques, nowadays widespread across all fields, can also be applied to the wealth of information derived from student interactions with serious games. Use of data science techniques can greatly improve the evaluation of games, and allow both teachers and institutions to make evidence-based decisions. This can increase both teacher and institutional confidence regarding the use of serious games in formal education, greatly raising their attractiveness. This paper presents a systematic literature review on how authors have applied data science techniques on game analytics data and learning analytics data from serious games to determine: (1) the purposes for which data science has been applied to game learning analytics data, (2) which algorithms or analysis techniques are commonly used, (3) which stakeholders have been chosen to benefit from this information and (4) which results and conclusions have been drawn from these applications. Based on the categories established after the mapping and the findings of the review, we discuss the limitations of the studies analyzed and propose recommendations for future research in this field.

Full publication

Computers & Education 141 (2019) 103612

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Applications of data science to game learning analytics data: A systematic literature review

Cristina Alonso-Fernández*, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón

Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, Madrid, Spain

ARTICLE INFO

Keywords:
Data science applications in education
Evaluation methodologies
Games
Teaching/learning strategies

ABSTRACT

Data science techniques, nowadays widespread across all fields, can also be applied to the wealth of information derived from student interactions with serious games. Use of data science techniques can greatly improve the evaluation of games, and allow both teachers and institutions to make evidence-based decisions. This can increase both teacher and institutional confidence regarding the use of serious games in formal education, greatly raising their attractiveness. This paper presents a systematic literature review on how authors have applied data science techniques on game analytics data and learning analytics data from serious games to determine: (1) the purposes for which data science has been applied to game learning analytics data, (2) which algorithms or analysis techniques are commonly used, (3) which stakeholders have been chosen to benefit from this information and (4) which results and conclusions have been drawn from these applications. Based on the categories established after the mapping and the findings of the review, we discuss the limitations of the studies analyzed and propose recommendations for future research in this field.

1. Introduction

Use of data science (e.g. artificial intelligence) techniques has spread over many fields, with a wide range of purposes. The huge, and constantly growing, amounts of data being captured allow complex techniques to provide insights which are potentially deeper than those that can be found by applying only traditional, and often simpler, methods.

The application of data science techniques perfectly fits interactive environments, where multiple data can be generated. One of these environments that allows for multiple interactions is games. In particular, the use of games with purposes beyond entertainment (e.g. learning, raising awareness or changing attitudes and behaviors), that is, so called *serious games* (SGs), has also increased in the last years. These types of games are especially popular in domains such as medicine or the military, and have proven their effectiveness for children and adolescents, as the familiarity of these users with gaming environments and the characteristics of games (interactivity, motivation, engagement) facilitate their interactions with serious games.

The collection and analysis of data has reached a great number of fields: in education, the fields of educational data mining (EDM) and learning analytics (LA), sometimes used interchangeably, are widely spread (Long & Siemens, 2011; Romero & Ventura, 2010). Their aim is to understand learners and their environments and improve the learning process through analysis of the data collected from students' interactions with the learning environment. As with any other highly interactive system, a lot of data can also be

* Corresponding author.
E-mail address: calonsofernandez@ucm.es (C. Alonso-Fernández).

https://doi.org/10.1016/j.compedu.2019.103612
Received 12 April 2019; Received in revised form 20 June 2019; Accepted 21 June 2019
Available online 24 June 2019
0360-1315/

gathered from serious games to guide data-based decision-making (Loh, Sheng, & Ifenthaler, 2015a, 2015b). Building up from the fields of educational data mining and learning analytics, which focus in education in general, game learning analytics (GLA) is defined as the collection, analysis and extraction of information from data collected from serious games (Alonso-Fernández et al., 2019, Alonso-Fernández, Pérez-Colado, Freire, Martínez-Ortiz, & Fernández-Manjón, 2019; Freire et al., 2016a, 2016b; Owen & Baker, 2019).

. The aim of the current paper is to conduct a systematic literature review on the applications of data science techniques to analyze game analytics data and/or learning analytics data from serious games. The rest of the paper is structured as follows: Section 2 provides a summary on related work; Section 3 describes the methodology used for the systematic literature review; Section 4 presents the results obtained; finally, Section 5 discusses the results, and presents the limitations and conclusions of the review.

2. Related work

The fields of serious games, learning analytics and data science have attracted considerable interest and attention in the last decade. While many works have been published related to these topics, we have not found any existing systematic literature reviews that combine the three topics together. The present work seeks to bridge this gap. In this section, we briefly present related literature reviews that involve at least one of the fields of serious games, learning analytics and data science; we also describe several that combine two of these topics.

We have found several literature reviews that examine serious games, each focusing on different aspects. That of (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012) focuses on the potential positive impact of gaming with respect to learning, skill enhancement and engagements, finding that the most frequently occurring outcomes and impacts were knowledge acquisition/content understanding, and affective and motivational outcomes. The review by (Petri & Gresse von Wangenheim, 2017) focused on the evaluation of serious games, finding that there are few approaches to systematically evaluate educational games. Special issues have also been published for different related fields like game visual analytics (Wallner, Canossa, & El-Nasr, 2018) or learning assessment (Berta & Moreno-Ger, 2018).

The possible applications of data science for games have also been studied. For instance, the work of (Yannakakis & Togelius, 2018) presents the major application areas of artificial intelligence methods within games: game-playing, content generation and player modeling. Regarding the possible applications of learning analytics on serious games, the literature review presented in the work of (Liu, Kang, Liu, Zou, & Hodson, 2017) focuses on uses of LA for assessment, but differs from our work in that we also focus on the specific data science techniques used and consider a broader set of purposes – not only assessment. Their results showed that SGs had a positive impact on learning and highlighted the importance of game design.

1. Although there are some similarities between the works described above and the systematic literature review presented in this paper, our work is different in that it focuses on serious games and the application of data science algorithms to game analytics data and/or learning analytics data coming from these types of games.

3. Method

3.1. Research questions

The main goal of this systematic literature review is to explore the applications of data science to game analytics data and/or learning analytics data from serious games. For this purpose, we have stated the following main research questions:

RQ1. What are the purposes for which data science has been applied to game analytics data and/or learning analytics data from serious games?
RQ2. What data science algorithms or techniques have been applied to game analytics data and/or learning analytics data from serious games?
RQ3. What stakeholders are the intended recipients of the analysis results?
RQ4. What results and conclusions have been drawn from these applications?

Additionally, we intend to extract some further information from the studies, to complement the results:

- The main purpose of the games (e.g. teaching, change behavior) and their domain (e.g. biology, math)
- The sample size of the studies, and the educational level of their participants
- The general characteristics of the in-game interaction data collected, and the data format used

3.2. Data collection

We follow a standard systematic literature review methodology, using a fixed set of queries on a pre-identified list of bibliographical databases, and clear inclusion/exclusion criteria.

3.2.1. Databases searched

We have queried 9 databases, including some of the main databases for education, computer science, and general science research. Specifically, we have searched: Association for Computing Machinery (ACM), Cambridge Journals Online, Education Resources Information Center (ERIC), IEEE Computer Society Digital Library (CDSL), IngentaConnect, Oxford University Press (journals), Science Direct, Scopus and Springer.

3.2.2. Search terms

To perform the searches on the databases, we focus on our three main terms of interest: data science, game analytics and learning analytics, and games. As seen in the Introduction, the terms "learning analytics" and "educational data mining" are sometimes used interchangeably, so we conducted two parallel searches, one for each of these terms. All searches are restricted to title, abstract and keywords.

- **Search query for game analytics**

We included the term "game analytics" and several alternative terms for "data science" and specific analysis techniques. Final search query:

("game analytics") AND ("artificial intelligence" OR "data mining" OR "machine learning" OR "data analysis" OR "deep learning")

- **Search query with learning analytics**

We included the terms "learning analytics", "games", and alternative terms for analysis techniques. Final search query:

("learning analytics") AND ("games") AND ("artificial intelligence" OR "data mining" OR "machine learning" OR "data analysis" OR "deep learning")

- **Search query with educational data mining**

As the term "educational data mining" includes the analysis of the data, we do not include additional terms of data analytics, and therefore used:

("educational data mining") AND ("games")

- **Additional search query on journal of artificial intelligence in education**

Finally, from our previous research, we encountered a specific journal on artificial intelligence in education, the International Journal of Artificial Intelligence in Education (Springer). We performed an additional search for papers of this journal which included the term "games".

3.2.3. Study selection

After removing duplicates, we scanned the title and abstract of all papers, comparing them against the inclusion and exclusion criteria below. After this first scanning, studies were classified as either *possible* or *excluded*. Clearly irrelevant publications were excluded, while those classified as *possible* were read (conclusions or even full text) to ensure relevancy. We examined possibly-relevant papers with our research questions in mind, to ensure that they provided enough information about purposes, techniques and stakeholders regarding the application of data science techniques to GLA data from serious games. Additionally, we looked for studies which included information on the games from which data was collected, sample sizes, and details on the interaction data collected from games; although this information was not mandatory for papers to be included in the review. No time restrictions were set.

Inclusion criteria

- Journals, conference papers or book chapters
- Include empirical evidence relating the outcomes of applying data science techniques to game analytics data and/or learning analytics data from serious games

Exclusion criteria

- Publications whose full text is not available
- Publications not written in English

3.3. Data analysis

For each of the studies selected for the literature review, we collected data on each of our research questions and conducted a mapping study to categorize the results of each research question. When available, we also collected additional information that

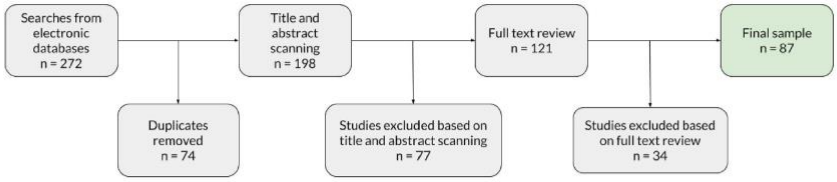

Fig. 1. Studies selection process for the systematic literature review.

could provide a more in-depth review of the applications of data science techniques to GLA data from SGs. We classified the data of the selected studies according to the following criteria:

- The main purposes of the analysis of the data collected from SGs (addressing RQ1)
- The algorithms or techniques used to analyze the data collected from SGs (addressing RQ2)
- The stakeholder that is the main beneficiary of the extracted information (addressing RQ3)
- The results and conclusions of the analysis of the data collected from SGs (addressing RQ4)
- The purpose of the serious game
- The domain of the serious game
- The sample size of the study
- The educational level or specific characteristics of the participants
- The in-game interaction data captured
- The format of the in-game interaction data captured

4. Results

4.1. Studies identified by search terms

Studies were retrieved in December 2018 using the search terms. In this first search, 272 studies were found. After analyzing the results, we excluded 74 duplicate publications, yielding 198 unique studies.

4.2. Studies selected using inclusion criteria

The selection process began with 198 studies. After scanning their titles and abstract, 77 studies were excluded for not meeting our inclusion criteria. An additional full text review was performed to ensure the suitability of the papers. On this final review, 34 studies were excluded for not meeting one or more of the inclusion criteria, such as not mentioning serious games or not collecting any data from the games. The final sample consists of 87 studies. Fig. 1 summarizes the full selection process.

Table 1 shows the total number of studies identified in the search process and meeting inclusion criteria from each database considered.

Regarding the year of publication of the selected studies, Fig. 2 shows the number of papers selected for the literature review for each year of publication. Note that two papers included in the review, to be published in 2019, were available online when the search was conducted; in the figure they are therefore considered as published in 2018. The figure shows an increased interest in the topics of the review from 2011 onwards.

4.3. Main purpose of studies

This subsection responds to RQ1 based on the studies that met all inclusion criteria:

RQ1. What are the purposes for which data science has been applied to game analytics data and/or learning analytics data from serious games?

After considering all qualifying studies, we mapped their main data science application purpose into one of 5 categories: learning assessment, study of in-game behaviors, game design or evaluation, student profiling, and interventions. Some studies focus on more than one of the previous purposes. Since several studies proposed frameworks to simplify the application of GLA data for SGs in specific contexts, but their primary purpose was unrelated to data science, we framed those studies under an additional, 6th category named framework proposals. Table 2 details the purposes of each data science related category, and lists two example studies for each.

Fig. 3 shows the main purpose of the selected studies. Assessment (32 studies, 36.8%) followed by in-game behaviors (27 studies, 31%), are the two main purposes.

Table 1
Number of studies identified in search and meeting inclusion criteria from each database.

Database	Number of studies identified in search	Number of studies meeting inclusion criteria
ACM	10	4
Cambridge	2	0
ERIC	32	12
IEEE CDSL	8	4
IngentaConnect	4	1
Oxford	0	0
Science Direct	8	5
Scopus	96	28
Springer	112	33
Total	272	87

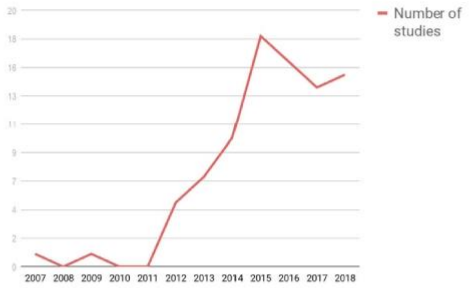

Fig. 2. Number of selected studies found per year of publication.

Table 2
Categories of purposes of data analysis application for the selected studies, with definition and two example studies.

Purpose category	Definition of purpose	Example studies
Assessment	Assess learning, predict performance	(Ke & Shute, 2015; R. S.; Baker, Clarke-Midura, & Ocumpaugh, 2016)
In-game behaviors	Study in-game players behaviors (e.g. persistence, engagement)	(Dicerbo, 2013; Kang, Liu, & Qu, 2017)
Game design	Validate game design	(Cano, Fernández-Manjón, & García-Tejedor, 2018; Tlili, Essalmi, Jemni, & Kinshuk, 2016)
Student profiling	Stablish categories of players profiles, differentiate players characteristics	(Denden, Tlili, Essalmi, & Jemni, 2018; Loh & Sheng, 2014)
Interventions	Study effect of in-game interventions (e.g. feedback messages, notification of performance)	(DeFalco et al., 2018; McCarthy, Johnson, Likens, Martin, & McNamara, 2017)
Framework proposals	Propose a framework for specific contexts	(Halverson & Owen, 2014; Nguyen, Gardner, & Sheridan, 2018)

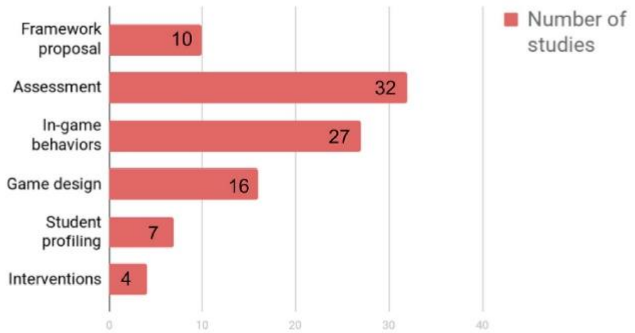

Fig. 3. Main purpose of data analysis in the selected studies. Some studies focus on several purposes.

Table 3
Data science techniques used in the studies, and number of papers that use each technique.

Data science technique	Number of papers using the technique
Supervised models	31
Linear/logistic regression	18
Regression/decision trees	7
Bayesian networks	6
Neural networks	4
Naïve Bayes	3
Bayesian knowledge tracing	3
Support vector machines	2
Unsupervised models	**35**
Correlation	17
Clustering	16
Factor analysis	2
Visualization	**36**
Performance metrics	15
Gameplay pathways	7
Use of in-game tools	5
Learning curves	4
Heatmaps of interactions	2

4.4. Data science algorithms or techniques

This subsection provides answers to RQ2:

RQ2. What data science algorithms or techniques have been applied to game analytics data and/or learning analytics data from serious games?

The data science algorithms and techniques used in the reviewed studies can be grouped into three main categories:

- Supervised algorithms: linear and logistic regression, regression and decision trees, support vector machines, Bayesian networks, neural networks, naive Bayes, and Bayesian knowledge tracing.
- Unsupervised algorithms: correlation, clustering, factor analysis.
- Visualization techniques: display of gameplay pathways, performance metrics, learning curves[1], heatmaps of interactions, use of in-game tools (frequency or duration).

Note that some studies present results of the application of various techniques and algorithms. Table 3 summarizes the techniques used and the number of studies that use each technique. The three main categories are used in a similar number of studies. For each of the three categories, we have specified the methods that are used in more than one study. We can see that linear models are the most used methods for supervised models (in 18 studies), while correlation and cluster analysis are the most widely used unsupervised methods in the studies (in 17 and 16 studies, respectively). Among the visualizations presented in the studies, a majority (15 studies) focus on displaying performance information.

4.5. Stakeholders

This subsection addresses RQ3:

RQ3. What stakeholders are the intended recipients of the analysis results?

The five stakeholders considered in the studies are: teachers/educators, serious game designers/developers, students/learners, researchers (or studies with research purposes) and parents. Fig. 4 shows the number of studies that focus on each of the stakeholders. Game designers or game developers are the main target of most studies (39 studies), closely followed by researchers, or studies with research purposes (37 studies); and teachers or educators (25 studies). We notice that when researchers are the main target of studies, they usually have additional roles (for instance, they also act as game designers or developers).

4.6. General information of the studies

Before moving to RQ4, this subsection looks at additional information of the studies regarding the serious games used, the participants and sample sizes, and the nature and contents of the captured interaction data.

4.6.1. Serious games used

Regarding the purpose of the serious games used, 55 studies (63.2%) use games that aim to teach, 8 (9.2%) to train and 6 (6.9%)

[1] Plot of speed or accuracy over game levels (Martin, Koedinger, Mitrovic, & Mathan, 2005).

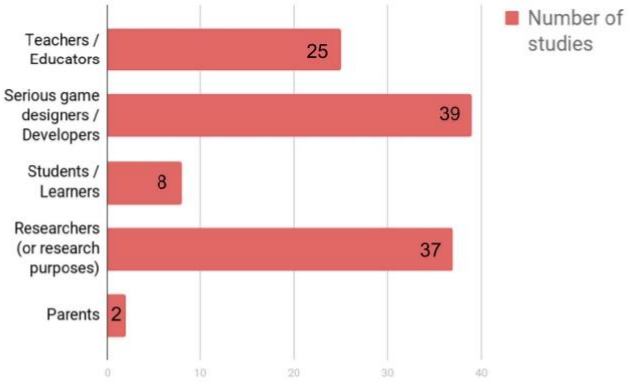

Fig. 4. Main stakeholder recipient of the analysis results in selected studies.

to assess. 2 studies use games to raise awareness and 1 to change behaviors. The remaining 15 studies (17.2%) do not clearly state the purpose of the games used.

The domain that most serious games in the selected studies focus on is mathematics (20 studies). 10 studies use science games and 4 studies focus on problem solving. Biology and physics are the domain of games in 3 studies each, while computer architecture, military, memory, language/reading and ability to design scientific investigations are the focus of games in 2 studies each. Other domains mentioned only in one study each include: business, programming, project management, ecology, research methodology, algorithmic and critical thinking, strategy planning or team work.

4.6.2. Participants and sample size

Regarding samples (N) used on the studies: 28 studies (32%) used fewer than 100 participants, while 27 studies (31%) use between 100 and 1000 participants. 7 studies (8%) used more than 1000 participants and only one study reported more than 10000 participants. The remaining 24 studies (27%) did not report data about participants. The samples of the 63 studies who reported participants are skewed by the highest values (mean = 1643 participants with SD = 10200; however, median = 116). Excluding outlier values from the calculations, that is, studies with over 1000 participants, the mean sample size for the remaining 55 studies drops to 161 participants (SD = 191). The high value of the standard deviation can be explained by the large number of papers with less than 100 participants (28 of these remaining 55 studies).

Analyzing the educational level of participants, 18 studies (20.7%) focus on primary school (up to 12 years old), 22 (25.3%) on secondary school (from 12 to 18 years old), 11 (12.6%) on undergraduates, 2 (2.3%) on graduates, and 8 (9.1%) on adults. 26 studies (30%) did not state this information.

4.6.3. Interaction data captured

An additional goal was to characterize data collected from the serious games. The game analytics data and/or learning analytics data captured in the selected studies include: completion times (in 30 studies); actions/interactions in general (28 studies); scores (14); correct/incorrect answers (11); clicks (9); attempts/tries (8); choices, errors/mistakes, answers (7); start/end, variables, completion, duration (4); events, performance, action sequences, contents accessed (3); phase/level changes, posts, location, context, success, items used/collected, progress, number of players (2). Other studies mention preferences, health, or use of in-game hints.

The format of the collected data is not specified in most studies: only 1 study reported the use of CSV as a format, 3 studies reported use of Experience API (xAPI) (ADL, 2012), 3 used XML, 3 used ad-hoc strings, and 6 studies used tables. The remaining 71 out of the 87 selected studies (81.6%) did not report any specific format of the collected data.

4.7. Results and conclusions of data analysis

This subsection addresses RQ4:

RQ4. What results and conclusions have been drawn from these applications?

The results and conclusions of the studies have been grouped based on the topics they are related to.

4.7.1. Results on assessment and student profiling

Several studies focus on assessment and learning predictions, also relating these with learners' characteristics and in-game behaviors. These results correspond to studies with purposes tagged in RQ1 as assessment, in-game behaviors and student profiling. A summary of the results of these studies is that:

GLA data can accurately predict games' impact:

- The application of GLA data can be useful both at real-time (*online*) and after the intervention is completed (*offline*) (Wiemeyer, Kickmeier-Rust, & Steiner, 2016), and for all stakeholders (Alonso-Fernández et al., 2019, Alonso-Fernández, Pérez-Colado, Freire, Martínez-Ortiz, & Fernández-Manjón, 2019). The analysis of interaction data can provide a means to measure the proven positive impact of games (Kosmas, Ioannou, & Retalis, 2018; Mavridis, Katmada, & Tsiatsos, 2017). However, most data is still captured after the game (Smith, Blackmore, & Nesbitt, 2015). Authors also point out the need for specific game learning analytics (Freire et al., 2016a, 2016b), or so-called serious games analytics (Loh et al., 2015a, 2015b), that differ from games analytics in general.
- Learning predictors: predictions of player success can often be accurately created based on log data (R. S. Baker et al., 2016), as the achievement system built into games may not be the most informative indicator of learning (Heeter, Lee, Medler, & Magerko, 2013). Some papers point out that their best predictors for measuring learning are based on the analysis of the player's exploration strategies (Horn et al., 2016; Kang et al., 2017; Käser, Hallinen, & Schwartz, 2017; Owen, Anton, & Baker, 2016; Smith, Hickmott, Southgate, Bille, & Stephens, 2016), or on the characteristics of player failures (Halverson & Owen, 2014). One study explored the relation between learning and students' facial behavior (Z. Xu & Woodruff, 2017). One study found, analyzing interactions in an online discussion forum, that the content that best explained and predicted learning was related to uncertainty, decision-making, time, collaboration and communication (Hernández-Lara, Perera-Lluna, & Serradell-López, 2019). In crowd-sourced serious games, three game-play metrics (active users, session counts and session time) were found to be good indicators of productivity by (Tellioglu, Xie, Rohrer, & Prince, 2014), while team cohesion and psychological safety may be good performance indicators in multiplayer serious games (Mayer, van Dierendonck, van Ruijven, & Wenzler, 2014). Also, behaviors such as avoiding a concept indicated poor performance (Ketamo, 2013). Implicit learning can also be adequately measured through behaviors (Rowe et al., 2017) and game log data (Rowe, Asbell-clarke, & Baker, 2015). Learning curves provide insights into learning (Peddycord-Liu et al., 2018) and can be studied for speed and accuracy (Eagle, 2009; R. S. J. D.; Baker, Habgood, Ainsworth, & Corbett, 2007).
- Recommendations to improve predictions: feature engineering improves performance of models with simple raw data (Owen & Baker, 2018). Additional information, such as the domain structure and the weights of competencies, improves accuracy of prediction models (Kickmeier-Rust, 2018). Exploratory data analysis (DiCerbo et al., 2015) and dynamical analysis (Snow, Allen, & McNamara, 2015) can uncover unexpected patterns and provide richer information about students' interactions.
- Creating assessment conditions: additional information may be extracted by letting teachers define assessment rules based on and combining generic game trace variables to obtain new information (Steiner, Kickmeier-Rus, & Albert, 2015). Complex assessment conditions can be created by combining some of the basic sets of traces (Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2014).

Performance is related to players' characteristics:

- Clusters of players in performance groups: players can be clustered into performance groups based on in-game actions (Martin et al., 2015; Slimani, Elouaai, Elaachak, Yedri, & Bouhorma, 2018; Freitas & Gibson, 2014; Forsyth et al., 2012; Lazo, Anareta, Duremdes, & Red, 2018; Polyak, von Davier, & Peterschmidt, 2017; Martinez-Garza & Clark, 2017; Chung, 2015) and in-game choices (Cutumisu, Blair, Chin, & Schwartz, 2017), which are also related to prior knowledge (Martinez-Garza & Clark, 2017). Some studies explore methods to differentiate experts from novice users (Loh and Sheng 2014, 2015a, 2015b). Once students are classified in a performance group, scores can be inferred when time or action sequences are added to the analysis (Gibson & Clarke-Midura, 2015). Tactics that lead to success can also be discovered with cluster analysis (Sharples & Domingue, 2016).
- Importance of understanding learners' characteristics: students with different learning characteristics may exhibit different learning behaviors (Liu, Lee, Kang, & Liu, 2016), for instance, different exploration strategies (Martin et al., 2013) or, in some cases, their age and gender (Wallner & Kriglstein, 2015). It is key to model students for effective adaptive instruction (Koedinger, McLaughlin, & Stamper, 2012), for instance, self-regulated learners tend to make better use of in-game curricular resources and may be more deliberate in their actions (Sabourin, Shores, Mott, & Lester, 2013) and high-performance students tend to use tools more appropriately (Liu, Kang, Lee, Winzeler, & Liu, 2015). Behaviors also depend on student background (Jaccard, Hulaas, & Dumont, 2017). Personalities can be identified based on actions and in-game choices (Denden et al., 2018).

Further information can be extracted from GLA data:

- Real-time information to stakeholders: in (Elaachak, Belahbibe, & Bouhorma, 2015) a system is presented that combines information for teachers, displaying a pie chart of students' performance, and for students, with assistance messages displayed on the screen according to their performance and progress. Some studies also provide systems that allow parents to receive real-time information about their children's learning (Ketamo, 2015; Roberts, Chung, & Parks, 2016).
- Measure other students' characteristics: some studies include additional applications of analysis of GLA data to track students' progress (Gweon et al., 2015), assess persistence (Dicerbo, 2013) or detect engagement (Ghergulescu & Muntean, 2016).

4.7.2. Results on serious game design

Several studies focus on applications to obtain further insight and improve serious game design and implementation. These results correspond to studies with purposes tagged in RQ1 as game design, interventions, and framework proposal. Studies have drawn

several conclusions and pointed out recommendations for serious game design based on findings of the analyzed game learning analytics data, including:

GLA data can validate serious game design:

- Several studies use game learning analytics data to validate serious game design (Cano et al., 2018; Harpstead, MacLellan, Aleven, & Myers, 2015; Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2012; Tlili et al., 2016), specific design choices (Cheng, Rosenheck, Lin, & Klopfer, 2017) and even to create new game mechanics (Ninaus, Kiili, Siegler, & Moeller, 2017) or to automatically discover speech act categories for dialogue-based educational games (Rus, Moldovan, Niraula, & Graesser, 2012).

Assessment can and should be integrated in serious game design:

- Recommendations for creating assessment in SGs: assessment design and learning context/task design should be considered in the early phase of game development (Ke & Shute, 2015). One study proposes a design approach to integrate data-driven assessment in game design (Ke, Shute, Clark, & Erlebacher, 2019). Debriefing via visualizations can improve understanding of outcomes (De Troyer, Helalouch, & Debruyne, 2016).
- Data to be collected: before applying GLA, and as part of the game design, it is highly recommended to specify and determine the game traces that will be collected (Tlili et al., 2016; Serrano-Laguna et al., 2018).
- Teachable agents: in educational games, teachable agents can help achieve deeper levels of learning that transfer outside the game (Pareto, 2014) and have a significant impact on in-game performance, preferably when designed to have low self-efficacy (Tärning, Silvervarg, Gulz, & Haake, 2018).

Importance of serious games characteristics:

- Difficulty: for games without adaptive difficulty, it is especially important to present a smooth difficulty curve. For (Hicks et al., 2016), a high difficulty level increased dropout. It is also important to classify students and modulate difficulty (Martinez-Garza & Clark, 2017). Allowing players to return to game areas with lower difficulty significantly reduced error rates and increased learning rate according to (Käser et al., 2013).
- Engagement and motivation: it is important to design for engagement by matching challenges with incentives and motivating activities (Pareto, 2014). Games should include motivational elements (Tlili et al., 2016). Engagement seems to decrease in internet experiments (Stamper et al., 2012).
- Feedback and interventions during play: too many metacognitive prompts during play may be detrimental (McCarthy et al., 2017). Mixed results obtained for self-efficacy enhancing interventions based on interaction-based affect detectors to enhance outcomes (DeFalco et al., 2018).

Identified challenges when designing serious games:

- Designing games for assessment: assessment routines are usually black boxes that teachers cannot inspect. Studies have identified a need for transparent and reliable assessment in educational games, based on assessment models that are ideally valid, easy to use, and provide meaningful educational information, while giving game industry evidence on game quality (Steiner et al., 2015). Different design decisions may be considered to explore how they affect learning outcomes (Plass et al., 2013). Adaptivity is also desired, although there is still a lack of real applications due to its high costs (Streicher & Smeddinck, 2016).
- Designing games with tracking features: game manufacturers are resistant to include data recording of learning evidence in their games, as they think it will increase costs and hamper the entertainment that encourages consumers to buy their games (Pereira, De Souza, & De Menezes, 2016).

Proposed frameworks:

- Some frameworks have been proposed to simplify tasks in serious game design, including: two game analytics frameworks for people with intellectual disabilities (García-Tejedor, Cano, & Fernández-Manjón, 2016; Nguyen et al., 2018), a game-based assessment model (Halverson & Owen, 2014), a framework to integrate design of event-stream features for analysis (Owen & Baker, 2018), a framework to support tracking and analysis of learners in-game activities (Hauge et al., 2014), a framework to help designers model experts' solving process almost automatically (Muratet, Yessad, & Carron, 2016), an interoperable adaptivity framework (Streicher & Roller, 2017), a framework for internet-scale experiments to inform and be informed by classroom and lab experiments (Stamper et al., 2012), an open-source SGs framework for sustainability (Y. Xu, Johnson, Lee, Moore, & Brewer, 2014) and a framework for a mobile game application for adults with cystic fibrosis (Vagg et al., 2018).

5. Discussion

We have found that most studies focus on assessment and learners' behaviors (RQ1). This suggests that, having established that games are indeed a useful tool for purposes beyond entertainment, there is an interest in analyzing interaction data to measure how much impact serious games have on players (mainly focusing on learning), and how that impact relates to players' in-game behaviors.

From our analysis of the methods used (RQ2), we found that visualization, supervised and unsupervised techniques are present in a similar number of papers. Among the data science techniques, the most widely used are linear models, correlations, and cluster techniques. All these methods are classical techniques, which may be surprising as newer, more complex and powerful techniques, in particular neural networks, are experiencing an important surge in popularity. A reason that may explain this result is the need of further evidence on how to widely and reliably apply these new complex techniques and the common difficulty to explain the results obtained, which has opened the debate about explainable AI (XAI) (Adadi & Berrada, 2018).

The main stakeholders considered, in a similar number of studies, are game designers/developers, and researchers, followed by teachers/educators (RQ3). This suggests that the analysis of data from games is used for several purposes including research, improving or validating game design, and providing information when applying games in educational scenarios. Although students only appear in 8 of the papers as the main direct stakeholders to benefit from the results of data analysis, they are always indirect recipients of the results, as the research, improvement and adaptivity of games and assessment techniques will make the use of games more effective and efficient for the ones who actually play the games, that is, the students/learners.

From the general information of the studies, we have found that most of the games used in the studies teach science-related topics, in particular mathematics. This result shows the intention to benefit from games' advantages to improve learning in a subject typically considered difficult for children and adolescents, and aligns with previous research which found that mathematics and science were the main areas for games that target primary education (Hainey, Connolly, Boyle, Wilson, & Razak, 2016). This may also be related with the fact that these domains have a clearly defined underlying model that simplifies assessment.

Sample sizes used in the studies are, in general, quite low (32% of the studies used less than 100 participants). This can potentially restrict the significance and generalization of their results, as well as the application of more complex algorithms such as deep neural networks, which require large amounts of data points to be adequately applied. The low sample size used in experiments is an important issue already pointed out by authors (Petri & Gresse von Wangenheim, 2017). Most participants were from primary and secondary schools, which aligns with the fact that the most commonly used games aim to teach mathematics.

Data collected from students' interactions included mainly completion times, actions or interactions in general, and scores. All these are common information that can be collected from any game but are, however, basic data that do not take full advantage of the rich interactions produced in games, as described in works on game analytics in entertainment games (Seif El-Nasr, Drachen, & Canossa, 2013). As some studies pointed out, the data to be collected is best identified at early stages of the game development, to ensure that it provides information with educational value. Most papers did not report the format in which they collected the data, so that we cannot know if they were using a standard or relying on their own data-formats. The latest scenario is less desirable, as it restricts the open sharing of the data for other purposes and requires an extra effort to replicate results with other techniques (Serrano-Laguna, Manero, Freire, & Fernández-Manjón, 2017). We have not found reports of any open data set of game analytics data or learning analytics data from serious games; this hinders research in this area, as testing out new data science techniques requires not only choosing the techniques themselves, but also developing a serious game and performing the experiments to collect its interaction data.

The analysis of GLA data from serious games has yielded, as expected, wide and varied results (RQ4). We can, however, extract some general findings from the conclusions and discussions of the studies analyzed:

- **Predicting games impact with GLA data**: raw data can be used to accurately predict impact (e.g. learning), including simple values from interactions (e.g. completion times, scores) but also more complex information such as kind of failures or exploration strategies. Adding information of the context is also recommended, as it can improve the models' accuracy. Also, the choice of data to analyze should ideally be taken during game design, to ensure that as much educationally-relevant data as possible is actually captured.
- **Importance of student profiling**: performance appears to be highly related to students' characteristics and behaviors, so it is recommended to create students' profiles or clusters to improve learning, including targeted feedback and adaptive learning experiences. The need to fit users' needs has also led authors to propose user-specific frameworks (e.g. for users with intellectual disabilities).
- **Designing serious games for assessment**: assessment needs to be formally and reliably integrated in the development phase of SGs to provide meaningful educational information. This should not damage costs or entertainment, as games need to maintain engaging and motivation features, while controlling for an adequate difficulty. GLA data can then be used to validate the game design and assessment.

5.1. Limitations

As in any other literature review, our work is limited by the search terms used. To try to minimize this limitation, we have included similar and commonly interchanged terms as well as an additional search on a specific journal. However, there still can be some relevant works that have not matched our search criteria.

5.2. Conclusions

This paper presents a systematic literature review on the applications of data science to game analytics data and/or learning analytics data collected from serious games. The goal of this literature review was to find how data science techniques have been used with interaction data from serious games, as we consider that the application of data science can increase the still-limited application

of serious games in education. Games have proven to be beneficial for learning in different domains, including more authentic learning and higher student engagement, both thanks to their greater degree of interactivity and immersion. Their limited application can be partly explained by a simple cost-benefit analysis: games are certainly costlier and more complex than other contents, and their advantages (for example, in terms of evaluating learning) are hard enough to measure that stakeholders may not be convinced of their overall return on investment (ROI). We consider that the information extracted from the application of data science techniques to interaction data from serious games can both reduce costs and complexity by simplifying game design and development, and measuring games' actual impact so the benefits of applying games are clearer for all stakeholders.

On this systematic literature review, we have identified 87 papers that reported evidence of the outcomes of the analysis of game analytics data and/or learning analytics data collected from serious games. We have classified them according to their specific purposes, methods of analysis used, stakeholders that benefit from the information and conclusions drawn from the analysis. These classifications can be used as a baseline for further research related to analysis of data from serious games.

Despite the diversity of the studies, we have been able to extract some notable common points and conclusions. The main purpose when analyzing data from serious games is assessment, most commonly with linear prediction models, simple correlation or cluster techniques, or visually displaying performance information. Learning predictions obtained are quite accurate and may be improved with some of the previous recommendations. The importance of student profiling as well as recommendations for integrating assessment into early phases of game design and development also stand out among the conclusions of the studies. Further research in this field should also bear in mind these recommendations to effectively and accurately assess students playing serious games.

Considering the studies presented in this review, we encourage researchers to consider large-enough sample sizes to ensure significant conclusions, and to decide in advance which data is to be collected from the games. In this sense, as a baseline, typical data such as completion times, interactions, or scores can and should be included; but research can benefit from moving on to more complex data extracted from in-game interactions. Regarding algorithms, we encourage researchers to compare classical techniques with new more complex ones (e.g. neural networks), to determine which ones draw the best results in each case. Finally, authors have pointed out a clear need for specific game learning analytics (GLA), where the use of standards to collect GLA data is desirable, as it allows the creation of open data sets in standard formats, such as xAPI (Serrano-Laguna et al., 2017), for research purposes, and simplifies results reproducibility and improvement, as well as testing of new techniques and integration of analytics as a module of larger systems.

Acknowledgements

This work has been partially funded by Regional Government of Madrid (eMadrid P2018/TCS4307), by the Ministry of Education (TIN2017-89238-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676, Erasmus + IMPRESS 2017-1-NL01-KA203-035259).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

ADL. (2012). *Experience API.* Retrieved March 20, 2016, from https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/.

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019a). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior, 99*, 301–309. https://doi.org/10.1016/j.chb.2019.05.036.

Berta, R., & Moreno-Ger, P. (2018). Introduction: Intelligent learning assessment in serious games. *International Journal of Serious Games, 5*(1)https://doi.org/10.17083/ijsg.v5i1.237.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686. https://doi.org/10.1016/j.compedu.2012.03.004.

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016a). Game learning analytics: Learning analytics for serious games. *Learning, Design, and Technology*, 1–29. https://doi.org/10.1007/978-3-319-17727-4_21-1.

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., & Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education, 102*, 202–223. https://doi.org/10.1016/j.compedu.2016.09.001.

Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning analytics as an assessment tool in serious games: A review of literature. *Serious games and edutainment applications* (pp. 537–563). . https://doi.org/10.1007/978-3-319-51645-5_24.

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015a). In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.). *Serious games analytics*https://doi.org/10.1007/978-3-319-05834-4.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 31–40.

Martin, B., Koedinger, K., Mitrovic, A., & Mathan, S. (2005). On using learning curves to evaluate ITS. *International Conference on Artificial Intelligence in Education*, 419–426.

Owen, E., & Baker, R. (2019). *Learning analytics for serious games.* (February). Retrieved from http://www.galanoe.eu/index.php/home/365-learning-analytics-for-serious-games.

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education, 107*, 68–90. https://doi.org/10.1016/j.compedu.2017.01.004.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532.

Seif El-Nasr, M., Drachen, A., & Canossa, A. (2013). In A. Drachen, & A. Canossa (Eds.). *Game analytics (M. Seif el-Nasr*https://doi.org/10.1007/978-1-4471-4769-5.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces, 50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014.

Wallner, G., Canossa, A., & El-Nasr, M. S. (2018). Introduction to the special issue on visual game analytics. *Information Visualization, 17*(3), 181–182. https://doi.org/10.1177/1473871617722040.

Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games.* https://doi.org/10.1007/978-3-319-63519-4.

Coded papers

Alonso-Fernández, C., Pérez-Colado, I., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019b). Improving serious games analyzing learning analytics data: Lessons learned. *Games and learning alliance: 7th international conference, GALA 2018, Palermo, Italy, December 5–7, 2018, proceedings: Vol. 10653*, (pp. 287–296). . https://doi.org/10.1007/978-3-030-11548-7_27.

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning, 32*(3), 267–280. https://doi.org/10.1111/jcal.12128.

Baker, R. S. J. D., Habgood, M. P. J., Ainsworth, S. E., & Corbett, A. T. (2007). Modeling the acquisition of fluent skill in educational action games. *User modeling 2007* (pp. 17–26). . https://doi.org/10.1007/978-3-540-73078-1_5.

Cano, A. R., Fernández-Manjón, B., & García-Tejedor, Á. J. (2018). Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. *British Journal of Educational Technology, 49*(4), 659–672. https://doi.org/10.1111/bjet.12632.

Cheng, M.-T., Rosenheck, L., Lin, C.-Y., & Klopfer, E. (2017). Analyzing gameplay data to inform feedback loops in the Radix Endeavor. *Computers & Education, 111*, 60–73. https://doi.org/10.1016/j.compedu.2017.03.015.

Chung, G. K. W. K. (2015). Guidelines for the design and implementation of game telemetry for serious games analytics. *Serious games analytics* (pp. 59–79). . https://doi.org/10.1007/978-3-319-05834-4_3.

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2017). Assessing whether students seek constructive criticism: The design of an automated feedback system for a graphic design task. *International Journal of Artificial Intelligence in Education, 27*(3), 419–447. https://doi.org/10.1007/s40593-016-0137-5.

De Troyer, O., Helalouch, A., & Debruyne, C. (2016). Towards computer-supported self-debriefing of a serious game against cyber bullying. In J. Dias, P. A. Santos, & R. C. Veltkamp (Eds.). *Games and learning alliance: 6th international conference, GALA 2017, Lisbon, Portugal, December 5–7, 2017, proceedings* (pp. 374–384). . https://doi.org/10.1007/978-3-319-50182-6_34.

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., et al. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education, 28*(2), 152–193. https://doi.org/10.1007/s40593-017-0152-1.

Denden, M., Tlili, A., Essalmi, F., & Jemni, M. (2018). Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors. *Smart Learning Environments, 5*(1), 1–19. https://doi.org/10.1186/s40561-018-0078-6.

Dicerbo, K. E. (2013). Game-based assessment of persistence. *Educational Technology & Society, 17*(1), 17–28.

DiCerbo, K. E., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., et al. (2015). An application of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.). *Serious games analytics* (pp. 319–342). . https://doi.org/10.1007/978-3-319-05834-4_14.

Eagle, M. (2009). Level Up : A frame work for the design and evaluation of educational games. *Artificial Intelligence, 434*(1), 339–341. https://doi.org/10.5507/bp.2011.016.

Elaachak, L., Belahbibe, A., & Bouhorma, M. (2015). Towards a system of guidance, assistance and learning analytics based on multi agent system Applied on serious games. *International Journal of Electrical and Computer Engineering (IJECE) Journal, 5*(2), 2088–8708. Retrieved from http://iaesjournal.com/online/index.php/IJECE.

Forsyth, C., Pavlik, P., Graesser, A., Cai, Z., Germany, M.-L., Millis, K., et al. (2012). Learning gains for core concepts in a serious game on scientific reasoning. *Proceedings of the 5th international conference on educational data mining, 1–4*. Retrieved from http://w.optimallearning.org/people/Articles/edm2012_short_2.pdf.

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016b). Game learning analytics: Learning analytics for serious games. *Learning, Design, and Technology*, 1–29. https://doi.org/10.1007/978-3-319-17727-4_21-1.

Freitas, S. de, & Gibson, D. (2014). Exploratory learning analytics methods from three case studies. Rhetoric and Reality: Critical Perspectives on Educational Technology. *Proceedings of Ascilite Dunedin, 2014*, 383–388.

García-Tejedor, Á. J., Cano, A. R., & Fernández-Manjón, B. (2016). GLAID: Designing a game learning analytics model to analyze the learning process in users with intellectual disabilities. *6th EAI international conference on serious games, interaction and simulation*. Retrieved from http://sgamesconf.org/2016/show/technical-session.

Ghergulescu, I., & Muntean, C. H. (2016). ToTCompute: A novel EEG-based TimeOnTask threshold computation mechanism for engagement modelling and monitoring. *International Journal of Artificial Intelligence in Education, 26*(3), 821–854. https://doi.org/10.1007/s40593-016-0111-2.

Gibson, D., & Clarke-Midura, J. (2015). *Some psychometric and design implications of game-based learning analytics. E-Learning systems, environments and approaches.* CELDA247–261. https://doi.org/10.1007/978-3-319-05825-2_17.

Gweon, G.-H., Lee, H.-S., Dorsey, C., Tinker, R., Finzer, W., & Damelin, D. (2015). Tracking student progress in a game-like learning environment with a Monte Carlo Bayesian knowledge tracing model. *Proceedings of the fifth international conference on learning analytics and knowledge - LAK '15* (pp. 166–170). . https://doi.org/10.1145/2723576.2723608.

Halverson, R., & Owen, V. E. (2014). Game-based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology, 9*(2), 111. https://doi.org/10.1504/ijlt.2014.064489.

Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2015). Replay analysis in open-ended educational games. *Serious games analytics* (pp. 381–399). . https://doi.org/10.1007/978-3-319-05834-4_17.

Hauge, J. B., Berta, R., Fiucci, G., Manjon, B. F., Padron-Napoles, C., Westra, W., et al. (2014). Implications of learning analytics for serious game design. *2014 IEEE 14th international conference on advanced learning technologies* (pp. 230–232). . https://doi.org/10.1109/ICALT.2014.73.

Heeter, C., Lee, Y.-H., Medler, B., & Magerko, B. (2013). Conceptually meaningful metrics: Inferring optimal challenge and mindset from gameplay. *Game analytics* (pp. 731–762). . https://doi.org/10.1007/978-1-4471-4769-5_32.

Hernández-Lara, A. B., Perera-Lluna, A., & Serradell-López, E. (2019). Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior, 92*, 600–612. https://doi.org/10.1016/j.chb.2018.03.001.

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016). Using game analytics to evaluate puzzle design and level progression in a serious game. *Proceedings of the sixth international conference on learning analytics & knowledge - LAK '16* (pp. 440–448). . https://doi.org/10.1145/2883851.2883953.

Horn, B., Hoover, A. K., Barnes, J., Folajimi, Y., Smith, G., & Harteveld, C. (2016). Opening the black box of play. *Proceedings of the 2016 annual symposium on computer-human interaction in play - CHI PLAY '16* (pp. 142–153). . https://doi.org/10.1145/2967934.2968109.

Jaccard, D., Hulaas, J., & Dumont, A. (2017). In J. Dias, P. A. Santos, & R. C. Veltkamp (Eds.). *Using comparative behavior analysis to improve the impact of serious games on students' learning experience*https://doi.org/10.1007/978-3-319-71940-5.

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757–770. https://doi.org/10.1016/j.chb.2016.09.062.

Käser, T., Busetto, A. G., Solenthaler, B., Baschera, G. M., Kohn, J., Kucian, K., et al. (2013). Modelling and optimizing mathematics learning in children. *International Journal of Artificial Intelligence in Education, 23*(1–4), 115–135. https://doi.org/10.1007/s40593-013-0003-7.

Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017). Modeling exploration strategies to predict student performance within a learning environment and beyond. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17* (pp. 31–40). . https://doi.org/10.1145/3027385.3027422.

Ke, F., & Shute, V. J. (2015). Serious games analytics. *Serious Games Analytics*. (January) https://doi.org/10.1007/978-3-319-05834-4.

Ke, F., Shute, V., Clark, K. M., & Erlebacher, G. (2019). *Interdisciplinary design of game-based learning platforms.* https://doi.org/10.1007/978-3-030-04339-1.

Ketamo, H. (2013). Agents and analytics - a framework for educational data mining with games based learning. *Proceedings of the 5th international conference on agents and artificial intelligence* (pp. 377–382). . https://doi.org/10.5220/0004331403770382.

Ketamo, H. (2015). User-generated character behaviors in educational games. *Healthcare Informatics Research, 21*, 57–68. https://doi.org/10.1007/978-981-287-408-5_5.

Kickmeier-Rust, M. D. (2018). Predicting learning performance in serious games. In S. Göbel,, A. Garcia-Agundez,, T. Tregel,, M. Ma,, J. Baalsrud Hauge,, & M. Oliveira, (Eds.). *Serious games* (pp. 133–144). . https://doi.org/10.1007/978-3-030-02762-9_14.

Koedinger, K., McLaughlin, E., & Stamper, J. (2012). Automated student model improvement. *Proceedings of the 5th international conference on educational data mining*

(pp. 17–24). . ISBN: 978-1-74210-276-4. Available at: http://educationaldatamining.org/EDM2012/uploads/procs/EDM_2012_proceedings.pdf.
Kosmas, P., Ioannou, A., & Retalis, S. (2018). Moving bodies to moving minds: A study of the use of motion-based games in special education. *TechTrends, 62*(6), 594–601. https://doi.org/10.1007/s11528-018-0294-5.
Lazo, P. P. L., Anareta, C. L. Q., Duremdes, J. B. T., & Red, E. R. (2018). Classification of public elementary students' game play patterns in a digital game-based learning system with pedagogical agent. *Proceedings of the 6th international conference on information and education technology - ICIET '18* (pp. 75–80). . https://doi.org/10.1145/3178158.3178160.
Liu, M., Kang, J., Lee, J., Winzeler, E., & Liu, S. (2015). Examining through visualization what tools learners access as they play a serious game for middle school science. *Serious games analytics* (pp. 181–208). . https://doi.org/10.1007/978-3-319-05834-4_8.
Liu, M., Lee, J., Kang, J., & Liu, S. (2016). What we can learn from the data: A multiple-case study examining behavior patterns by students with different characteristics in using a serious game. *Technology, Knowledge and Learning, 21*(1), 33–57. https://doi.org/10.1007/s10758-015-9263-7.
Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330. https://doi.org/10.1016/j.chb.2014.07.022.
Loh, C. S., & Sheng, Y. (2015a). Measuring expert performance for serious games analytics: From data to insights. *Serious games analytics* (pp. 101–134). . https://doi.org/10.1007/978-3-319-05834-4_5.
Loh, C. S., & Sheng, Y. (2015b). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*(1), 5–19. https://doi.org/10.1007/s10639-013-9263-y.
Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015b). Serious games analytics: Theoretical framework. *Serious games analytics* (pp. 3–29). . https://doi.org/10.1007/978-3-319-05834-4_1.
Martin, T., Aghababyan, A., Pfaffman, J., Olsen, J., Baker, S., Janisiewicz, P., et al. (2013). Nanogenetic learning analytics. *Proceedings of the third international conference on learning analytics and knowledge - LAK '13* (pp. 165). . https://doi.org/10.1145/2460296.2460328.
Martinez-Garza, M. M., & Clark, D. B. (2017). Investigating epistemic stances in game play with data mining. *International Journal of Gaming and Computer-Mediated Simulations, 9*(3), 1–40. https://doi.org/10.4018/ijgcms.2017070101.
Martin, T., Petrick Smith, C., Forsgren, N., Aghababyan, A., Janisiewicz, P., & Baker, S. (2015). Learning fractions by splitting: Using learning analytics to illuminate the development of mathematical understanding. *The Journal of the Learning Sciences, 24*(4), 593–637. https://doi.org/10.1080/10508406.2015.1078244.
Mavridis, A., Katmada, A., & Tsiatsos, T. (2017). Impact of online flexible games on students' attitude towards mathematics. *Educational Technology Research & Development, 65*(6), 1451–1470. https://doi.org/10.1007/s11423-017-9522-5.
Mayer, I., van Dierendonck, D., van Ruijven, T., & Wenzler, I. (2014). Stealth assessment of teams in a digital game environment. *Lecture Notes in Computer Science, 8605*, 224–235. https://doi.org/10.1007/978-3-319-12157-4_18.
McCarthy, K. S., Johnson, A. M., Likens, A. D., Martin, Z., & McNamara, D. S. (2017). *Metacognitive prompt overdose: Positive and negative effects of prompts in iSTART. Grantee submission.* 404–405 Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED577125&site=ehost-live.
Muratet, M., Yessad, A., & Carron, T. (2016). Understanding learners' behaviors in serious games. In F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, & R. W. H. Lau (Eds.). *Advances in web-based learning - ICWL 2015* (pp. 195–205). . https://doi.org/10.1007/978-3-319-47440-3_22.
Nguyen, A., Gardner, L. A., & Sheridan, D. (2018). A framework for applying learning analytics in serious games for people with intellectual disabilities. *British Journal of Educational Technology, 49*(4), 673–689. https://doi.org/10.1111/bjet.12625.
Ninaus, M., Kiili, K., Siegler, R. S., & Moeller, K. (2017). Data-driven design decisions to improve game-based learning of fractions. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 10653*LNCS. (pp. 3–13) https://doi.org/10.1007/978-3-319-71940-5_1.
Owen, V. E., Anton, G., & Baker, R. (2016). Modeling user exploration and boundary testing in digital learning games. *Proceedings of the 2016 conference on user modeling adaptation and personalization - UMAP '16* (pp. 301–302). . https://doi.org/10.1145/2930238.2930271.
Owen, V. E., & Baker, R. S. (2018). Fueling prediction of player decisions: Foundations of feature engineering for Optimized behavior modeling in serious games. *Technology, Knowledge and Learning*0123456789. https://doi.org/10.1007/s10758-018-9393-9.
Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. *International Journal of Artificial Intelligence in Education, 24*(3), 251–283. https://doi.org/10.1007/s40593-014-0018-8.
Peddycord-Liu, Z., Harred, R., Karamarkovich, S., Barnes, T., Lynch, C., & Rutherford, T. (2018). In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.). *Learning curve analysis in a large-scale, drill-and-practice serious math game: Where is learning support needed?* (pp. 436–449). . https://doi.org/10.1007/978-3-319-93843-1_32.
Pereira, H. A., De Souza, A. F., & De Menezes, C. S. (2016). A computational architecture for learning analytics in game-based learning. *Proceedings - IEEE 16th international conference on advanced learning technologies, ICALT 2016* (pp. 191–193). . https://doi.org/10.1109/ICALT.2016.3.
Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., et al. (2013). Metrics in simulations and games for learning. *Game analytics* (pp. 697–729). . https://doi.org/10.1007/978-1-4471-4769-5_31.
Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology, 8*(NOV), 1–16. https://doi.org/10.3389/fpsyg.2017.02029.
Roberts, J. D., Chung, G. K. W. K., & Parks, C. B. (2016). Supporting children's progress through the PBS KIDS learning analytics platform. *Journal of Children and Media, 10*(2), 257–266. https://doi.org/10.1080/17482798.2016.1140489.
Rowe, E., Asbell-clarke, J., & Baker, R. S. (2015). Serious games analytics to measure implicit science learning. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.). *Serious games analytics* (pp. 343–360). . https://doi.org/10.1007/978-3-319-05834-4.
Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A. G., Barnes, T. M., et al. (2017). Assessing implicit science learning in digital games. *Computers in Human Behavior, 76*, 617–630. https://doi.org/10.1016/j.chb.2017.03.043.
Rus, V., Moldovan, C., Niraula, N., & Graesser, A. (2012). Automated discovery of speech act categories in educational games. *Proceedings of International Conference on Educational Data Mining*, 25–32.
Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education, 23*(1–4), 94–114. https://doi.org/10.1007/s40593-013-0004-6.
Serrano-Laguna, Á., Manero, B., Freire, M., & Fernández-Manjón, B. (2018). A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes. *Multimedia Tools and Applications, 77*(2), 2849–2871. https://doi.org/10.1007/s11042-017-4467-6.
Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a little for big improvements: Application of learning analytics and videogames for student assessment. *Procedia Computer Science, 15*, 203–209 Elsevier.
Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2014). Application of learning analytics in educational videogames. *Entertainment Computing, 5*(4), 313–322. https://doi.org/10.1016/j.entcom.2014.02.003.
Sharples, M., & Domingue, J. (2016). *Adaptive and adaptable learning.* Switzerland: Lecture Notes in Computer Science13–16. 9891 https://doi.org/10.1007/978-3-319-45153-4.
Slimani, A., Elouaai, F., Elaachak, L., Yedri, O. B., & Bouhorma, M. (2018). Learning analytics through serious games: Data mining algorithms for performance measurement and improvement purposes. *International Journal of Emerging Technologies in Learning, 13*(1), 46–64. https://doi.org/10.3991/ijet.v13i01.7518.
Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A meta-analysis of data collection in serious games research. *Serious games analytics* (pp. 31–55). . https://doi.org/10.1007/978-3-319-05834-4_2.
Smith, S. P., Hickmott, D., Southgate, E., Bille, R., & Stephens, L. (2016). Exploring play-learners' analytics in a serious game for literacy improvement. In T. Marsh, M. Ma, M. F. Oliveira, J. Baalsrud Hauge, & S. Göbel (Eds.). *Serious games* (pp. 13–24). . https://doi.org/10.1007/978-3-319-45841-0_2.
Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). The dynamical analysis of log data within educational games. *Serious games analytics* (pp. 81–100). . https://doi.org/10.1007/978-3-319-05834-4_4.
Stamper, J. C., Lomas, D., Ching, D., Ritter, S., Koedinger, K. R., & Steinhart, J. (2012). The rise of the super experiment. *Proceedings of the 5th International Conference on Educational Data Mining*, 196–199. https://doi.org/10.1177/0003122412458508.

Steiner, C. M., Kickmeier-Rus, M. D., & Albert, D. (2015). Making sense of game-based user data: Learning analytics in applied games. *International conference E-learning* (pp. 195–198). . https://doi.org/10.1017/CBO9781107415324.004.

Streicher, A., & Roller, W. (2017). Interoperable adaptivity and learning analytics for serious games in image interpretation. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 10474*, (pp. 598–601). LNCS. https://doi.org/10.1007/978-3-319-66610-5_71.

Streicher, A., & Smeddinck, J. D. (2016). In R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, & K. Zweig (Eds.). *Personalized and adaptive serious games* (pp. 332–377). Springer. https://doi.org/10.1007/978-3-319-46152-6_14.

Tärning, B., Silvervarg, A., Gulz, A., & Haake, M. (2018). Instructing a teachable agent with low or high self-efficacy – does similarity attract? *International Journal of Artificial Intelligence in Education, 29*(1), 89–121. https://doi.org/10.1007/s40593-018-0167-2.

Tellioglu, U., Xie, G. G., Rohrer, J. P., & Prince, C. (2014). Whale of a crowd: Quantifying the effectiveness of crowd-sourced serious games. *2014 computer games: AI, animation, mobile, multimedia, educational and serious games (CGAMES)* (pp. 1–7). . https://doi.org/10.1109/CGames.2014.6934151.

Tlili, A., Essalmi, F., Jemni, M., & Kinshuk (2016). An educational game for teaching computer architecture: Evaluation using learning analytics. *2015 5th international conference on information and communication technology and accessibility, ICTA 2015*https://doi.org/10.1109/ICTA.2015.7426881.

Vagg, T., Tan, Y. Y., Shortt, C., Hickey, C., Plant, B. J., & Tabirca, S. (2018). MHealth and serious game analytics for cystic fibrosis adults. *2018 IEEE 31st international symposium on computer-based medical systems (CBMS), 2018-June* (pp. 100–105). . https://doi.org/10.1109/CBMS.2018.00025.

Wallner, G., & Kriglstein, S. (2015). Comparative visualization of player behavior for serious game analytics. *Serious games analytics* (pp. 159–179). . https://doi.org/10.1007/978-3-319-05834-4_7.

Wiemeyer, J., Kickmeier-Rust, M., & Steiner, C. M. (2016). Performance assessment in serious games. *Serious Games, 13*, 273–302. https://doi.org/10.1007/978-3-319-40612-1_10.

Xu, Y., Johnson, P. M., Lee, G. E., Moore, C. A., & Brewer, R. S. (2014). *Makahiki : An open source serious game framework for sustainability education and conservation, Vol. 8*. International Association for Development of the Information Society.

Xu, Z., & Woodruff, E. (2017). Person-centered approach to explore learner's emotionality in learning within a 3D narrative game. *Proceedings of the seventh international learning analytics & knowledge conference on - LAK '17* (pp. 439–443). . https://doi.org/10.1145/3027385.3027432.

6.1.2. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study

Full citation

Cristina Alonso-Fernández, Iván Martínez-Ortiz, Rafael Caballero, Manuel Freire, Baltasar Fernández-Manjón (2020): **Predicting students' knowledge after playing a serious game based on learning analytics data: A case study.** Journal of Computer Assisted Learning, vol. 36, no. 3, pp. 350-358, June 2020. DOI: 10.1111/jcal.12405.

Impact metrics: JCR 2019, Impact Factor: 2.126, Q2 in Education & Educational Research.

Abstract

Serious games have proven to be a powerful tool in education to engage, motivate, and help students learn. However, the change in student knowledge after playing games is usually measured with traditional (paper) prequestionnaires–postquestionnaires. We propose a combination of game learning analytics and datamining techniques to predict knowledge change based on in-game student interactions. We have tested this approach in a case study for which we have conducted preexperiments–postexperiments with 227 students playing a previously validated serious game on first aid techniques. We collected student interaction data while students played, using a game learning analytics infrastructure and the standard data format Experience API for Serious Games. After data collection, we developed and tested prediction models to determine whether knowledge, given as posttest results, can be accurately predicted. Additionally, we compared models both with and without pretest information to determine the importance of previous knowledge when predicting postgame knowledge. The high accuracy of the obtained prediction models suggests that serious games can be used not only to teach but also to measure knowledge acquisition after playing. This will simplify serious games application for educational settings and especially in the classroom easing teachers' evaluation tasks.

Full publication

Received: 5 June 2019 | Revised: 14 October 2019 | Accepted: 3 November 2019
DOI: 10.1111/jcal.12405

ARTICLE

Journal of Computer Assisted Learning WILEY

Predicting students' knowledge after playing a serious game based on learning analytics data: A case study

Cristina Alonso-Fernández | Iván Martínez-Ortiz | Rafael Caballero | Manuel Freire | Baltasar Fernández-Manjón

Computer Science Faculty, Complutense University of Madrid, Madrid, Spain

Correspondence
Cristina Alonso-Fernández, Computer Science Faculty, Complutense University of Madrid, Madrid, Spain.
Email: calonsofernandez@ucm.es

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12405.

Abstract

Serious games have proven to be a powerful tool in education to engage, motivate, and help students learn. However, the change in student knowledge after playing games is usually measured with traditional (paper) prequestionnaires–postquestionnaires. We propose a combination of game learning analytics and data mining techniques to predict knowledge change based on in-game student interactions. We have tested this approach in a case study for which we have conducted preexperiments–postexperiments with 227 students playing a previously validated serious game on first aid techniques. We collected student interaction data while students played, using a game learning analytics infrastructure and the standard data format Experience API for Serious Games. After data collection, we developed and tested prediction models to determine whether knowledge, given as posttest results, can be accurately predicted. Additionally, we compared models both with and without pretest information to determine the importance of previous knowledge when predicting postgame knowledge. The high accuracy of the obtained prediction models suggests that serious games can be used not only to teach but also to measure knowledge acquisition after playing. This will simplify serious games application for educational settings and especially in the classroom easing teachers' evaluation tasks.

KEYWORDS
learning analytics, serious games, game-based learning, assessment, e-learning, xAPI

1 | INTRODUCTION

Serious games (SGs) are games or game-like applications with purposes beyond entertainment (Michael & Chen, 2005). In education, SGs have proven to be an effective way to promote learning due to their engaging and immersive nature (Boyle, Connolly, Hainey, & Boyle, 2012), which increases students' participation in the learning process.

To adequately evaluate players' knowledge when using an SG, a common method is the use of two external questionnaires for each player, one *before* playing (pretest) and another *after* playing (posttest). This methodology is the most common and accepted practice in the medical domain to evaluate the efficacy of SGs (Calderón & Ruiz, 2015). However, several authors have pointed out that this external and summative evaluation of learning is error prone and reduces the time to play (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012; Frederick-Recascino, Liu, Doherty, Kring, & Liskey, 2013). Preexperiments–postexperiments can also be used to measure how students' knowledge improves when using an already evaluated game. This evaluation of acquired knowledge is the focus of this work. The goal of this research is to showcase the use of in-game interactions to predict students' knowledge using data mining techniques.

The goal of learning analytics (LA) techniques is to collect, analyse and report data to understand and optimize learners' contexts on educational systems (Long & Siemens, 2011). The high interactivity of SGs provides large quantities of interaction data, allowing the application of LA techniques. Game learning analytics (GLA) is defined as the

process of capturing, storing, analysing, and obtaining information from players' interactions with an SG (Freire et al., 2016).

We consider that in-game interactions (i.e., GLA) can be analysed to automatically and accurately determine users' knowledge after playing. This allows us to evaluate players as an integral part of the playing, avoiding disruption of the game experience and without needing an external measure. We propose to determine players' knowledge following these stages:

- Game validation phase: The SG is validated using the traditional and widely accepted prequestionnaires–postquestionnaires, while also tracking players' interactions. Then, different supervised machine learning models are tested to predict knowledge, taking as input the interaction data (and, optionally, the pretest) and validated against actual knowledge results (given in the posttest).
- Game deployment phase: Once a sufficiently accurate prediction model is obtained, in subsequent applications of the game, students' knowledge after playing can be automatically predicted on the basis of in-game interactions. This greatly simplifies the application of games in the classroom by no longer requiring students to fill prequestionnaires–postquestionnaires. The predicted students' knowledge can be used as an indicator for teachers to know how much students finally know about the topic covered in the game.

We test our approach by conducting a case study to determine if players' knowledge, as measured by a posttest, could be accurately predicted by applying machine learning techniques to previously gathered information (pretest and in-game interactions). We are also interested in determining the best prediction models and the most relevant information when predicting knowledge. Additionally, we want to know the extent to which availability of the pretest (which directly measures players' knowledge before the game) affects the accuracy of predictions.

The rest of the paper is structured as follows: Section 2 reviews the related work for data mining techniques applied for knowledge predictions in education; Section 3 states the research questions of the current case study; Section 4 describes the methodology, including participants, experimental design, and materials and instruments; Section 5 summarizes the best results obtained from the predictions models; Section 6 presents a discussion of the results in relation to the research questions; and Section 7 contains conclusions, limitations, and future work.

2 | RELATED WORK

Data mining techniques have been applied in education to understand students and their learning scenarios, in a discipline called educational data mining (EDM; Baker & Yacef, 2009). For instance, the U.S. Department of Education has studied the use of data mining techniques for different purposes such as prediction to enhance learning (Bienkowski, Feng, & Means, 2012). On page 28, they highlight that "inferring what a user knows […] requires looking at accumulated data that represents the interactions between students and the learning system." Data mining techniques have been applied to LA data to predict students' knowledge and prevent their failure, helping teachers and students to improve their teaching and learning processes (Shahiri, Husain, & Rashid, 2015). These predictions usually target performance, knowledge, score, or marks, either via regression analysis to find relationships between students' variables or via classification to group students (Romero & Ventura, 2010). Authors have also revised the models used to predict students' knowledge finding that they commonly include neural networks and decision trees (Shahiri et al., 2015), Bayesian networks, rule-based systems, regression, and correlation analysis (Romero & Ventura, 2010). The analysis of 240 EDM publications by (Peña-Ayala, 2014) yielded that student modelling and assessment are the main targets of EDM application, with predictive models for classification being the most frequently applied, in particular Bayes theorem, logistic regression, and decision trees. Support vector machine models have also been used to predict faculty performance evaluation, using different kernel methods (Deepak, Pooja, Jyothi, Kumar, & Kishore, 2016). On a recent literature review, we found out that assessment is commonly the target of the application of data mining techniques to LA data, specially applying classical techniques such as regression and decision trees (Alonso-Fernández, Calvo-Morata, Freire, Martinez-Ortiz, & Fernández-Manjón, 2019).

Despite the highly accurate results obtained by applying data mining to LA data, we have not found widely accepted approaches that systematically measure players' knowledge after using SGs. The instruments that measure knowledge (usually prequestionnaires–postquestionnaires) have to be developed ad hoc for each game (Petri & Gresse von Wangenheim, 2016). This is a significant limitation that influences a lack of replicability and of well-defined models (Petri & Gresse von Wangenheim, 2017). In the last years, data-based evaluations have been used to measure the learning process in a discipline called *stealth assessment* (V. Shute & Kim, 2014). Stealth assessment relies on capturing sequences of actions made by students while interacting with a highly interactive and immersive tool (e.g., a game), to obtain information of what students know and do not know at each moment (V. Shute & Ventura, 2013). This information is updated when new data are captured and is later used to evaluate students. This promising research line has high implementation cost yet, as solutions need to be developed ad hoc for each game. We consider that our two-step approach can be more easily generalized to a broader set of educational games. To the best of our knowledge, we have not found specific studies where data mining has been used to improve the evaluation of students' knowledge when using SGs that had already been experimentally validated, as we propose in our approach.

3 | RESEARCH QUESTIONS

To test our two-step approach, we must first verify that we can infer students' knowledge after playing an SG. This motivates the initial research question of this case study:

- Q1.1. Can we accurately predict student knowledge from previous knowledge and interactions with an SG? (we refer to this as the *pre + game* condition, because we use both pretest and in-game interactions to build the prediction models).
- In case we can predict it, our next step is to find the most accurate predictions models and the most relevant information for those predictions. Therefore, we also propose a follow-up research question:
- Q1.2. If we can indeed predict student knowledge after playing an SG, what prediction models perform best, and what are the most relevant variables for these models?

We also want to explore the possibility of predicting knowledge solely from in-game interactions, without relying on any pretest information. We call this the *game-only* condition. For this condition, we again look for the most suitable models and the most relevant information for predictions and will compare it against results from Q1.1 and Q1.2, which use the *pre + game* condition. Therefore, we propose the following additional research questions:

- Q2.1. Can we accurately predict student knowledge solely from interactions with an SG? (*game-only* condition)
- Q2.2. What are the best prediction models and the most valuable information towards those predictions?
- Q2.3. Is the *pre + game* condition (proposed in Q1.1) more effective at predicting student knowledge than the *game-only* condition (proposed in Q2.1)?

To answer these research questions, we have used a previously validated SG on a preexperiment–postexperiment with a control group (Marchiori et al., 2012). As the game was later updated to a new technology, to carry out the "Game validation phase" of our approach and build the prediction models, we first conducted a new set of experiments with prequestionnaires–postquestionnaires while tracking GLA data from players' interactions.

4 | METHODOLOGY

4.1 | Participants

The experiments for this case study involved N = 227 high school students from a charter school in Madrid, Spain. We conducted two sessions with 28 students as an initial formative evaluation. Their feedback helped us to test the remote data collection in the school settings and prepare for the main experience. Out of the remaining population (N = 199), gender was not obtained for 15 students due to an error handling a questionnaire. For the other 184 students, the gender distribution was 46.7% males and 53.3% females. The median age was 14 years old. Figure 1 summarizes the gender distribution by age. In terms of gender and age, this sample is representative of the student population in Madrid (Comunidad de Madrid, 2016; Instituto Nacional de Evaluación Educativa—Ministerio de Educación, 2017).

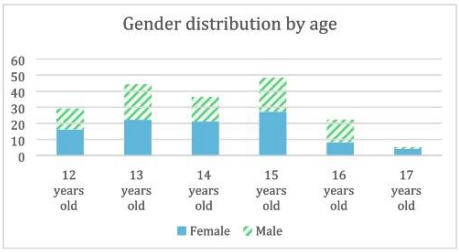

FIGURE 1 Gender distribution by age of the participants in the experiments

4.2 | Experimental design

At the beginning of each session, the teacher (playing the role of the session manager) gave each student a unique identification code that allowed them to access the game and which was used in all questionnaires instead of any personally identifying information (Perez-Colado et al., 2019). Then, each student/player completed, in this order: (a) a questionnaire before starting the game (pretest), (b) a complete game session in the chosen SG, and (c) a questionnaire after playing the game (posttest). Each player is therefore linked to the three data sources via the unique player identification code, which acts as a pseudonym and reduces potential privacy pitfalls. The complete experiment was designed to fit into a standard 50-min session. Students could repeat the game levels as many times as they wanted up to 30 min. Figure 2 summarizes the research design of the experiment. The experiment was reviewed and approved by the school management as an educational activity. Students were informed about the data capturing, and the school signed an informed consent.

4.3 | Materials and instruments

4.3.1 | The *First-Aid Game*

The *First-Aid Game* is a game-like simulation with narrative structure that aims to teach basic life-support manoeuvres for players for 12–16 years old, focusing on chest pain, unconsciousness, and

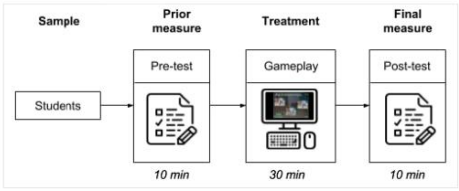

FIGURE 2 Research design of the experiment: All players completed a pretest, a game session, and a posttest

FIGURE 3 Textual and visual options available in the serious game *First-Aid Game*

choking. These three situations are depicted as game levels. In each scenario, players can interact with the main character or use a mobile phone (visible at the bottom right corner in the first screenshot of Figure 3) to call the emergency services. The game offers multiple-choice situations (second screenshot of Figure 3) that feature the specific first aid knowledge to be learnt through the game (e.g., Heimlich manoeuvre to avoid chocking). Players learn whether their decisions are appropriate or not: If they choose an incorrect answer, either the game reports the critical error and its consequences and lets them try again until they choose the correct answer or the game allows you to continue to later discover the consequences (and it is reflected in the final score). Options may be textual or visual (see Figure 3). The game includes random elements to improve reflection and replayability (e.g., availability of a semi-automatic external defibrillator). The three levels can be replayed as many times as desired during the available time. After each level is completed, a score is provided to indicate players whether their actions were mostly correct or not. This score does not directly measure players' knowledge but challenges them to replay levels where they made many mistakes.

The game was developed and evaluated by the e-UCM Research Group and actual emergency physicians (e-UCM, 2012; Marchiori et al., 2012). The game was validated in 2012 with an experiment that included pretests–posttests to measure players' knowledge and a control group to compare the game effect against that of a theoretical–practical demonstration by a trained instructor. Players in the experimental group gained, on average, 2.07 points on a 10-point scale, compared with control group learners who gained 3.61 points. This proved that the game achieves its goal of making player learn first aid procedures. The game was later adapted and updated to the Unity 3D videogame engine using *uAdventure* (Perez Colado, Perez Colado, Martínez-Ortiz, Freire, & Fernandez-Manjon, 2017), an authoring tool developed by the same group. This included the tracking of GLA data used in the present work.

4.3.2 | Questionnaires

As mentioned above, two questionnaires were used in the experiments. The pretest consisted of three parts: demographic variables (players' gender and age); a first aid knowledge questionnaire with 15 multiple-choice questions, also used in the original experiment to validate the game (Marchiori et al., 2012) and covering the game contents; and a game habits questionnaire with eleven 5-point Likert questions on game habits obtained from (Manero, Torrente, Freire, & Fernández-Manjón, 2016) and slightly adapted for this experiment. The posttest consisted of two parts: a repetition of the first aid knowledge questionnaire used in the pretest (to compare results) and a questionnaire to evaluate the experience itself, with five 5-point Likert questions assessing the experience, and optional free-text sections for feedback. The scores on the first aid knowledge questionnaires are defined as the total number of correct answers. Therefore, possible scores ranged from 0 to 15 points. Internal consistency of the scale used was ensured when the test was created in the original validation experiment (Marchiori et al., 2012). These questionnaires were a simplification of the ones used in the medical domain and had been previously validated.

4.3.3 | GLA data collected

During gameplays, a software component embedded in the SG (called a *tracker*) sent out players' interactions (i.e., traces) to an external server, developed by the e-UCM Research Group, using the Experience API (xAPI) standard's SG profile (Serrano-Laguna et al., 2017) to transmit and store interaction data.

The collected xAPI data[1] were analysed to derive variables that described how each player played the game. The specific information to be tracked from the game as well as the derived variables were chosen on the basis of the learning and game designs of the game as specified in its LA model (Perez-Colado, Alonso-Fernández, Freire-Moran, Martinez-Ortiz, & Fernández-Manjón, 2018) and in collaboration with domain experts. The process of capturing the data following the LA model for this and two other games is described in more detail in Alonso-Fernández et al. (2019). The variables extracted from the in-game interactions included whether the game has been completed or not, the first and maximum scores achieved in each of the three game levels, the number of times each level was repeated, the interactions

[1]The data that support the findings of this study are available from the corresponding author upon reasonable request.

with game elements, and whether specific questions were answered correctly or not. Appendix A provides the full list of variables derived from the xAPI statements and their detailed descriptions.

4.3.4 | Prediction models

All prediction models were built using RStudio and taking, as inputs, the complete set of variables derived from xAPI data, as described above. Models were created with and without pretest information as input, to further determine if the pretest is essential to predict players' knowledge after playing or not. The target variable of the predictions is the posttest score. Two types of models were created: linear models to predict exact score in range (0–15), and classification models to predict pass/fail category (establishing pass as 8 points out of 15).

We selected the algorithms most widely used in the literature for data mining applied to LA data: regression and decision trees, and linear and logistic regression. Although trees can show complex, non-linear relationships providing easy-to-understand models, regression is useful when data are not extremely complex or not a lot of data are gathered. Additionally, these models are white box models, which will allow us to relate the results obtained to our input data to obtain further information related to the traces collected from the game. A priori, our dataset is not too large, so regression should still be viable; however, if complex relationships appear, trees are expected to be better at discovering them. Different models were tested, including and excluding variables and interactions between variables. We additionally included two methods commonly mentioned in the literature: naïve Bayes for classification and support vector machines for regression (SVR), testing different non-linear kernels (polynomial, radial basis, and sigmoid; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997), and tuning the different parameters, with the ranges recommended in the literature (Hsu, Chang, & Lin, 2016). Models were compared using 10-fold cross-validation. When predicting pass/fail, and because data were not balanced (169 students passed the posttest, whereas only 30 failed it), classification models were created with an undersample of 78 students (40% from the fail class and 60% from the pass class) and tested on the original sample.

5 | RESULTS

We first verified that again in this case, study knowledge increase was significant. Pretest and posttest score variables were not normal (Shapiro–Wilk test yielded $p < .01$). Therefore, to measure knowledge change without assuming a normally distributed population, we use the paired sample Wilcoxon signed-rank test. The test showed a significant increase ($p < .05$, $r = -.41$) from pretest scores (mean = 8.06, $SD = 2.05$) to posttest scores (mean = 9.83, $SD = 2.38$). This proves replicability of results from the validation experiment and allows us to create predictive models.

As stated in the previous section, we used decision trees, logistic regression, and naïve Bayes classifier for pass/fail predictions and regression trees, linear regression, and SVR with non-linear kernels for score predictions. For pass/fail predictions, Table 1 provides precision, recall, and error, measured as misclassification rate. For score predictions, we provide the mean and the standard deviation of the error. Notice that the error is measured in the score scale of 0–15. Table 1 also summarizes the best models obtained, highlighting the best results in bold font.

The first rows of Table 1 summarize the best models when predicting pass/fail and posttest score with both pretest and in-game information (*pre + game* condition): Logistic regression provides the lowest misclassification rate and the highest recall when predicting pass/fail, whereas SVR provides the lowest mean error for score predictions (although the standard deviation is higher than for regression trees). The lower half of the table is dedicated to the *game-only* condition, summarizing results from models that predict pass/fail and posttest scores solely with in-game information. Logistic regression again provides the most accurate predictions of pass/fail, whereas SVR methods provide the lowest mean error when predicting score.

For score prediction models, 95% confidence intervals (CIs) for predictions were calculated using bootstrapping. The score scale of 0–15 was used, and then results were normalized to the 0–10 scale typically used for grading in Spain. In the *pre + game* condition, the regression tree obtained a mean posttest score prediction (in 0–10 scale) of 6.56 with 95% CI of [3.74, 8.53], whereas linear regression obtained a mean prediction of 6.62 with 95% CI of [4.76, 7.53]. In the *game-only* condition, the regression tree obtained a mean score

TABLE 1 Prediction models for posttest pass/fail and score, with and without pretest information

	Pass/fail prediction				Score prediction (scale [0–15])	
		Success measure		Error		
Pretest?	Data mining model	Precision	Recall	MR	Data mining model	Mean (*SD*)
Yes (*pre + game*)	Decision tree	81.6%	94.2%	16.2%	Regression tree	2.22 **(0.55)**
	Logistic regression	89.8%	**98.3%**	**10.5%**	Linear regression	1.68 (1.44)
	Naïve Bayes classifier	**92.6%**	89.7%	15.1%	SVR (non-linear kernels)	**1.47** (1.33)
No (*game-only*)	Decision tree	88.6%	92.4%	17.3%	Regression tree	2.38 **(0.62)**
	Logistic regression	87.2%	**98.8%**	**12.7%**	Linear regression	1.89 (1.54)
	Naïve Bayes classifier	**89.7%**	90.6%	16.9%	SVR (non-linear kernels)	**1.56** (1.37)

prediction of 6.54 with 95% CI of [4.06, 8.42], whereas linear regression obtained 6.55 with 95% CI of [4.6, 7.74]. These CI results confirm that linear regression models outperform regression tree models in terms of accuracy for score prediction.

Not all the variables used as input for the models (listed in Appendix A) were found to have the same relevance towards predictions. With pretest information, in predicting pass/fail results, the most relevant variables were the pretest score, the final game score, and the number of times each situation was repeated. In predicting score, the most relevant were the maximum score achieved in the "chest pain" game level, the number of interactions with the game character, and failure when answering the question regarding the Heimlich position. Solely with game interactions, most important variables to predict pass/fail include interactions with the game character and the first and maximum score achieved in the "chest pain" level. To predict score, the number of interactions with the game character appears as a relevant variable in all models, together with the maximum score in both the "chest pain" and "unconsciousness" levels, and failure on the "Heimlich position" question.

6 | DISCUSSION

In this section, we answer the research questions stated in Section 3, on the basis of discussing the results presented in Section 5.

Q1.1. Can we accurately predict student knowledge from previous knowledge and interactions with an SG? (*pre + game* condition)

Yes. The highly accurate results allow prediction of knowledge (as posttest results) from previous information (pretest and game interactions). As expected, more accurate predictions are obtained for pass/fail categories, but score predictions are still reasonably accurate. In many situations, as this case study not dealing with core subjects (e.g., math), it may suffice for teachers to know if students have acquired enough knowledge to pass or fail the subject. In fact, classification is most widely used for education (Peña-Ayala, 2014; Shahiri et al., 2015).

Q1.2. If we can indeed predict student knowledge after playing an SG, what prediction models perform best, and what are the most relevant variables for these models?

To classify players in pass/fail categories, the best model is a logistic regression, as naïve Bayes has higher precision but lower recall. This is not surprising, as we are predicting a binary variable, a task well suited for logistic regression (Maalouf, 2011), instead of classifying among several categories. To predict posttest scores, results are not as precise, but SVR yields the lowest error. As SVR was built with non-linear kernels, this result may be an indicator of non-linear relationships between the variables.

The most valuable variables for these predictions include the number of interactions with the game character, the final game score, and the maximum score achieved in the "chest pain" level, although some pretest variables were also slightly significant (e.g., pretest score). A possible explanation is that game mechanics and educational design relate scores in each level (and final score) to knowledge, low scores being a consequence of making domain-relevant errors. Although scores in this specific game are a good indicator of knowledge, this may not be the case for all games. For interactions with the game character, models show that higher number of interactions predicted lower scores. As in most situations, the game design forces players to retry when making an error, and the number of interactions will increase when errors are made, suggesting a "trial and error" strategy. This design decision may explain why the number of interactions is a good predictor of knowledge. If this mechanic appears in other games, a good predictor may be the number of retries or errors. Notice that this discussion is possible as we are analysing results of a white box model (logistic regression); for black box models, such a discussion, if possible, will not be as straightforward (Dreiseitl & Ohno-Machado, 2002).

Q2.1. Can we accurately predict student knowledge solely from interactions with an SG? (*game-only* condition)

Yes. We have obtained accurate prediction results for posttest scores solely from in-game interactions, without pretest information. More accurate results are obtained when predicting pass/fail classification, but still accurate results are obtained for score predictions.

Q2.2. What are the best prediction models and the most valuable information towards those predictions?

To predict pass/fail results, the best prediction model appears to be a logistic regression, as in the case of the *pre + game* condition. Although other models provide a slightly better precision, recall is higher and misclassification rate much lower than in the other models. To predict posttest scores, in the 0–15 range, the best prediction model is again based on SVR. However, we notice that the standard deviation is higher than in the decision tree (which has a higher mean error).

The most useful variables for these predictions again include the number of interactions with the game character and the first and maximum scores obtained in the "chest pain" game level. Regarding interactions with game character, the higher the number of interactions, the lower the score predicted, so the same discussion as above is valid. An unexpected finding is the greater relevance of (first and maximum) scores in the "chest pain" level compared with those of the two other levels. A possible explanation is that, although players could play levels in any order, the "chest pain" level appears in the left-most part of the screen, so most students, accustomed to scanning media left to right (Spalek & Hammad, 2005), played it first. Therefore, this result suggests that the first level students play may have a greater influence on knowledge acquisition.

Q2.3. Is the *pre + game* condition (proposed in Q1.1) more effective at predicting student knowledge than the *game-only* condition (proposed in Q2.1)?

Yes but only slightly. Models in the *pre + game* condition show better predictions in general, but, as shown in Table 1, models in the *game-only* condition still obtain accurate results. That is, results show similarly accurate predictions with and without pretest information, if in-game interactions remain as input for the models.

7 | CONCLUSIONS AND LIMITATIONS

This work presents a case study of an approach to measure students' knowledge after playing SGs based on their game interactions, after an initial priming phase to create the prediction models. We have tested the "Game validation phase" of our approach with an already validated game to ensure that the game indeed fulfils the goal of making players learn. The high accuracy of the models obtained on this case study show that we can indeed predict knowledge after playing, using both pretest and game LA data as inputs. From the models tested, we have seen that pretest information is relevant but by no means essential. Therefore, it is possible to infer students' knowledge solely from in-game interactions. Another option is to use the pretest as a formal evaluation of previous knowledge and, comparing it with the predictions of subsequent knowledge, calculate how much players have learned playing.

After the initial phase to formally validate the game and train the algorithms, games could be deployed automatically (with no posttest required), because the knowledge gained by playing can be inferred from game interactions, as described in the "Game deployment phase" of our approach. Using the most accurate prediction model, a prediction of knowledge after playing the game for each player could be obtained without the need to carry out the posttest. This prediction could be used for teacher as evaluation, allowing the use of games not only to teach but also to measure knowledge gained by players, while reducing costs of experiments in both time and effort. This approach also allows games to be played by larger samples of students, whose results can be automatically predicted. In some scenarios (e.g., in online games), after playing, students could optionally accept the score predicted from their interactions or take a real posttest. Another possibility for teachers is to use the games as an exam to evaluate students, taking the predicted knowledge as their score.

The encouraging results obtained on this case study suggest that our two-step approach proposed may be generalized at least to other similar cases, such as games for procedural learning or game-like simulations with narrative structure that are quite common in several domains (e.g., military and medicine). Both can provide similar interaction data, and therefore, by following the described steps, a similar approach could be applied. The specific data to be collected on each case should be driven by the specific educational design of each game, similarly to what we have done in our case study following the LA model (Perez-Colado et al., 2018), although we expect that relevant interactions for one type of game will also be relevant for similar games.

Based on our results, we can extract some lessons learned, which may be useful for SG designers and researchers that wish to assess players with SGs. For our game, we have found specific data (e.g., number of interactions with game elements) that seem to be related with knowledge and an emphasis on the results achieved on the early phases of the game (e.g., scores in game levels). In Section 6, we have provided possible explanations for these results linked with both the game mechanics and the educational game design. We therefore advise both mechanics and educational designs to be considered when deciding which interaction data to capture from SGs. Using an accepted standard format (e.g., xAPI-SG) is a clear recommendation, as it simplifies tracking, replicability of models, and integration in a wider range of systems.

This work has some limitations. The most relevant is that the SG used was evaluated in a previous preexperiment–postexperiment using an accepted existing measurement test on the topics of the game (Marchiori et al., 2012). This allowed us to apply prediction models, as it was proven that students learned with the game. Other limitation is that the data used are from one SG and a single school, which could potentially bias the results. However, we consider that the approach could be generalized for a wider range of games and students with similarly accurate results.

Future lines of work include testing these approaches with larger datasets and more complex games to attempt to replicate the highly encouraging results reported in this work. We also consider that better results could be obtained using games with knowledge prediction as an explicit design goal. Therefore, another line is testing this approach with games originally designed to be evaluated with these techniques, for instance, designing both the game and the interaction data to be gathered to improve score predictions. The predictions of learning obtained may also be used for players' assessment. We plan to study the use of games for assessment and propose a similar approach as the one described on this work focusing on the in-game stealth assessment of players (V. J. Shute & Moore, 2017). Although the most promising algorithms identified in the literature (Deepak et al., 2016; Peña-Ayala, 2014; Romero, López, Luna, & Ventura, 2013; Romero & Ventura, 2010; Shahiri et al., 2015) have been tested, some other more complex or even nontraditional methods could also be explored.

ORCID

Cristina Alonso-Fernández https://orcid.org/0000-0003-2965-3104

REFERENCES

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers in Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.10361

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301–309. https://doi.org/10.1016/j.chb.2019.05.036

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief* (pp. 1–57. Retrieved from). Washington, DC: SRI International. https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. https://doi.org/10.1016/j.chb.2011.11.020

Calderón, A., & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, *87*, 396–422. https://doi.org/10.1016/j.compedu.2015.07.011

Clark, D. B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., & Sengupta, P. (2012). Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov modeling. In *Assessment in Game-Based Learning* (pp. 173–199). https://doi.org/10.1007/978-1-4614-3546-4_10

Comunidad de Madrid. (2016). Datos y cifras de la Educación 2016/2017.

Deepak, E., Pooja, G. S., Jyothi, R. N. S., Kumar, S. V. P., & Kishore, K. V. (2016). SVM kernel based predictive analytics on faculty performance evaluation. *2016 International Conference on Inventive Computation Technologies (ICICT)*, 1–4. https://doi.org/10.1109/INVENTIVE.2016.7830062

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*(x), 155–161. https://doi.org/10.1.1.10.4845

e-UCM. (2012). First-Aid Game. Retrieved from http://first-aid-game.e-ucm.es/

Frederick-Recascino, C., Liu, D., Doherty, S., Kring, J., & Liskey, D. (2013). Articulating an experimental model for the study of game-based learning. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 8018 LNCS* (pp. 25–32). https://doi.org/10.1007/978-3-642-39226-9_4

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In *Learning, Design, and Technology* (pp. 1–29). https://doi.org/10.1007/978-3-319-17727-4_21-1

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2016). A practical guide to support vector classification. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Instituto Nacional de Evaluación Educativa—Ministerio de Educación, C. y D. (2017). Sistema Estatal de Indicadores de la Educación Edición 2016. Retrieved from http://www.mecd.gob.es/dctm/inee/indicadores/2016/17768–sistemaestatalindicadores2016-27-3-2017.pdf?documentId=0901e72b824643f0%5Cnhttp://www.mecd.gob.es/inee/sistema-indicadores/Edicion-2016.html

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 31–40.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

Manero, B., Torrente, J., Freire, M., & Fernández-Manjón, B. (2016). An instrument to build a gamer clustering framework according to gaming preferences and habits. *Computers in Human Behavior*, *62*, 353–363. https://doi.org/10.1016/j.chb.2016.03.085

Marchiori, E. J., Ferrer, G., Fernandez-Manjon, B., Povar-Marco, J., Suberviola, J. F., & Gimenez-Valverde, A. (2012). Video-game instruction in basic life support maneuvers. *Emergencias*, *24*(6), 433–437.

Michael, D. R., & Chen, S. L. (2005). Serious games: Games that educate, train, and inform. *Education*, October, *31*, 1–95. https://doi.org/10.1145/2465085.2465091

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Perez Colado, I., Perez Colado, V., Martínez-Ortiz, I., Freire, M., & Fernandez-Manjon, B. (2017). uAdventure: The eAdventure reboot—Combining the experience of commercial gaming tools and tailored educational tools. *IEEE Global Engineering Education Conference (EDUCON)*, 1754–1761. Retrieved from http://www.e-ucm.es/drafts/e-UCM_draft_304.pdf

Perez-Colado, I. J., Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Simva: Simplifying the scientific validation of serious games. *9th IEEE International Conference on Advanced Learning Technologies (ICALT)*.

Perez-Colado, I. J., Alonso-Fernández, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2018). Game learning analytics is not informagic! *IEEE Global Engineering Education Conference (EDUCON)*.

Petri, G., & Gresse von Wangenheim, C. (2016). How to evaluate educational games: A systematic literature review. *Journal of Universal Computer Science*, *22*(7), 992–1021. https://doi.org/10.3217/jucs-022-07-0992

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, *107*, 68–90. https://doi.org/10.1016/j.compedu.2017.01.004

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, *50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shute, V., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of research on educational communications and technology: Fourth edition* (pp. 311–321). https://doi.org/10.1007/978-1-4614-3185-5_3

Shute, V., & Ventura, M. (2013). Stealth assessment. In *The SAGE encyclopedia of educational technology* (p. 91). https://doi.org/10.4135/9781483346397.n278

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*.

Spalek, T. M., & Hammad, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science*, *16*(1), 15–18. https://doi.org/10.1111/j.0956-7976.2005.00774.x

How to cite this article: Alonso-Fernández C, Martínez-Ortiz I, Caballero R, Freire M, Fernández-Manjón B. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *J Comput Assist Learn*. 2019;1–9. https://doi.org/10.1111/jcal.12405

APPENDIX A: | LIST OF GAME INTERACTION VARIABLES

Table A1 provides the full list of variables derived from the xAPI-SG statements collected from students' gameplays. These variables are used as input for the prediction models. Table A1 provides the variables names, types, and detailed description.

TABLE A1 Variables selected from game interaction xAPI statements

Variable name	Type	Description
gameCompleted	Binary (true, false)	True if learner completed the game; false otherwise
Score	Numerical in range [0, 10]	Total score obtained in the game
maxScoreCP	Numerical in range [0, 10]	Maximum score obtained in "chest pain" level
maxScoreU	Numerical in range [0, 10]	Maximum score obtained in "unconsciousness" level
maxScoreCH	Numerical in range [0, 10]	Maximum score obtained in "choking" level
firstScoreCP	Numerical in range [0, 10]	First score obtained in "chest pain" level
firstScoreU	Numerical in range [0, 10]	First score obtained in "unconsciousness" level
firstScoreCH	Numerical in range [0, 10]	First score obtained in "choking" level
timesCP	Integer	Number of times student completed "chest pain" level
timesU	Integer	Number of times student completed "unconsciousness" level
timesCH	Integer	Number of times student completed "choking" level
int_patient	Integer	Number of interactions with patient (game character, NPC)
int_phone	Integer	Number of interactions with phone (game element)
int_saed	Integer	Number of interactions with defibrillator (game element)
failedEmergency	Binary (true, false)	True if learner failed, at least once, the question about the emergency number; false otherwise
failedThrusts	Binary (true, false)	True if learner failed, at least once, the question about the number of abdominal thrusts per minute; false otherwise
failedHName	Binary (true, false)	True if learner failed, at least once, the question about the name of Heimlich manoeuvre; false otherwise
failedHPosition	Binary (true, false)	True if learner failed, at least once, the question about the initial position for Heimlich manoeuvre; false otherwise
failedHHands	Binary (true, false)	True if learner failed, at least once, the question about the hand position for Heimlich manoeuvre; false otherwise

6.1.3. Evidence-based evaluation of a serious game to increase bullying awareness

Full citation

Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2020): **Evidence-based evaluation of a serious game to increase bullying awareness.** Interactive Learning Environments, 2020. DOI: 10.1080/10494820.2020.1799031.

Impact metrics: JCR 2019, Impact Factor: 1.938, Q2 in Education & Educational Research.

Abstract

Game Learning Analytics can be used to conduct evidence-based evaluations of the effect that serious games produce on their players by combining in-game user interactions and traditional evaluation methods. We illustrate this approach with a case-study where we conduct an evidence-based evaluation of a serious game's effectiveness to increase awareness of bullying. In this paper, we describe: (1) the full process of tracking in-game interactions, analyzing the traces collected using the standard xAPI-SG format, and deriving game learning analytics variables (to be used as evidences); and (2) the use of those variables to predict the increase in bullying awareness. We consider that this process can be generalized and replicated to systematize, and therefore simplify, evidence-based evaluations for other serious games based on the interaction data of their players.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2020.1799031

Routledge
Taylor & Francis Group

Check for updates

Evidence-based evaluation of a serious game to increase bullying awareness

Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz and Baltasar Fernández-Manjón

Department of Software Engineering and Artificial Intelligence, Faculty of Computer Science, Complutense University of Madrid, Madrid, Spain

ABSTRACT
Game Learning Analytics can be used to conduct evidence-based evaluations of the effect that serious games produce on their players by combining in-game user interactions and traditional evaluation methods. We illustrate this approach with a case-study where we conduct an evidence-based evaluation of a serious game's effectiveness to increase awareness of bullying. In this paper, we describe: (1) the full process of tracking in-game interactions, analyzing the traces collected using the standard xAPI-SG format, and deriving game learning analytics variables (to be used as *evidences*); and (2) the use of those variables to predict the increase in bullying awareness. We consider that this process can be generalized and replicated to systematize, and therefore simplify, evidence-based evaluations for other serious games based on the interaction data of their players.

ARTICLE HISTORY
Received 5 February 2020
Accepted 15 July 2020

KEYWORDS
Serious games; learning analytics; stealth assessment; technology-enhanced learning; game-based learning; e-learning

Introduction

Serious Games (SGs) are defined as games whose purposes go beyond simple entertainment, such as teaching, training, or changing perceptions. For instance, serious games have been used to raise awareness about social problems or to effectively change players' attitudes or behaviors (Peng et al., 2010; Xu et al., 2014).

The traditional methodology to evaluate serious games uses paired external formal questionnaires to assess players: one before the application of the game (pre-test) and one after the gameplay is finished (post-test). Results of both questionnaires are then compared to evaluate the game's effect on its players and, therefore, whether it achieves its intended goals (Calderón & Ruiz, 2015).

Given the availability of in-game interaction data, data-based approaches can gain deeper evaluation insights than those which only rely on paired questionnaires. Game Learning Analytics (GLA) brings together the fields of Learning Analytics (in education) and Game Analytics (in game industry), and is defined as the tracking, collection and analysis of data from the interaction of players with serious games for several purposes, such as improving the game design, understanding players' mental processes, or assessing their learning (Alonso-Fernández, Calvo-Morata, et al., 2019; Freire et al., 2016). On a previous work, we started to investigate how to combine both traditional formal and widely-accepted methods (pre-post experiments) and more recent and powerful techniques (Game Learning Analytics) to assess players' characteristics using a serious game (Alonso-Fernández et al., 2020). This latter approach poses the basis for evidence-based serious games evaluation.

CONTACT Cristina Alonso-Fernández calonsofernandez@ucm.es

This paper presents a case-study of an evidence-based evaluation of the effect of *Conectado* using GLA and data mining techniques. *Conectado* is a serious game to raise awareness about bullying and cyberbullying for players between 12 and 17 years old. The game is designed as a tool for teachers to use in their classrooms to spark discussions on this topic after all players/students have shared the common experience of playing, and has been validated through several experiments in schools (Calvo-Morata et al., 2019, 2020). The evidence-based evaluation of the effect of the game is performed in two steps: (1) **obtaining evidences from the serious game:** describes the collection of data from the interaction between players and games, including its representation (in our case, using the standard xAPI-SG Profile), and the processing and analysis of the resulting data to obtain relevant GLA variables; and (2) **using those evidences to predict the effect of the serious game:** the GLA variables derived from xAPI-SG traces are used as input for prediction models, which, once trained, can predict how the characteristics that the game is built to change (in our case, bullying awareness) will effectively change. This evidence-based evaluation aims to simplify players' assessment in further deployment of the games, where the effect on players could be calculated based solely on the actions they take in the game (*evidences*).

The rest of this paper is structured as follows: the "Related work" section presents previous research on assessment of learning for serious games; the "Materials and methods" section describes the game and methodology used in this case-study, including the experiments that we have carried out to collect interaction data from *Conectado*, the variables we derived from those traces, and the models we chose to predict increase of bullying awareness based on in-game actions. Prediction results are then presented in the "Results" section. Finally, the "Discussion" analyzes our process and its results section, while the "Conclusion" section summarizes the contributions of this work.

Related work

The application of serious games in different domains has increased the interest towards the integration of players' assessment in the videogame itself (Shoukry et al., 2014). The method to conduct the assessment of learning has been traditionally based on external measures (e.g. pre-post tests), but has recently been shifting towards more evidence-based approaches. For instance, *stealth assessment* (Shute et al., 2017; Shute & Ventura, 2013) aims to embed the assessment in the game, collecting evidences in a non-disruptive manner while the game is in play. These evidences are then used to update a game-based model that informs the results of players' assessment.

In a literature review on the uses of learning analytics data for assessment in serious games, Liu *et al* pointed out the need of more studies that combine different data sources, for example traditional measures, such as questionnaires, with more dynamic data, such as in-game metrics. They also noticed the lack of standard procedures to guide researchers on how to use in-game data (Liu et al., 2017). These conclusions have been pointed out by other researchers, highlighting the need of more research on how serious games can be effectively used for assessment, and which characteristics of such games contribute or detract to their validity for assessment (Kato & Klerk, 2017). On a previous literature review on data applications in serious games, we also found out that assessment was the main purpose of such applications, but that more research was required to establish general approaches, and that the sample sizes used in such studies should be increased (Alonso-Fernández, Calvo-Morata, et al., 2019).

Materials and methods

As previously mentioned, the game to be evaluated is *Conectado*, a serious game to raise awareness about bullying and cyberbullying. In *Conectado*, players play in first person as a student that transfers into a new school, and, during the first week, becomes increasingly bullied by classmates. Those aggressions happen both in the school and at home, where the bullying continues via social media (which makes it cyberbullying). The game has a linear flow and, depending on the actions

taken, such as mentioning the problem with the character's parents or teachers, players will reach one of the three different game endings. By design, player's choices only have an immediate effect on the next dialogs but do not affect the main storyline until just before the ending. This ensures that all players will go through all the situations represented in the game, while still experiencing their actions as meaningful, even while they have minimal effect on the overall flow of the game. Linear play also makes all playthroughs of comparable length, and provides all players with a common experience for their in-class post-game discussions.

This case-study with *Conectado* comprises two phases. In the *experiment phase*, we have collected data from pre-post questionnaires, together with *Conectado* in-game interaction data as represented using the xAPI-SG Profile. In the *analysis phase*, we have processed those xAPI-SG traces to derive a set of GLA variables, which we have then used as input for prediction models to predict the increase in bullying awareness. The actual increase in bullying awareness, to compare the predictions of the models with, is obtained by comparing the scores of each players' pre- and post- tests as gathered in the experiment phase.

Figure 1 depicts how this two-step methodology is organized. Pre-post questionnaires are collected first, through experiments with actual students in their classrooms, and the results of these experiments are later used to derive the target variable (increase in bullying awareness) of the prediction models (green lines). In-game interaction traces for *Conectado* are automatically collected using a game tracker, that is, a reusable component that allows developers to communicate the SG with an analytics platform. These traces are processed to obtain the values of the GLA variables (which have been previously defined based on knowledge of the game, or by relying on a default set of variables), and these variables are then used as input for the prediction models (red lines). We have used Python both to process the xAPI-SG traces and to train and use the prediction models.

Collecting interaction data from Conectado

The data used in the case-study was obtained from $N = 1109$ participants (ages 12–17) from 11 schools around Spain. In all experiments, participants completed a pre-test, a gameplay of *Conectado*, and a post-test, in that order. Minimal time elapsed between the gameplay and either of the two tests, and the complete sessions lasted a total of around 50 minutes, fitting in an average-length lecture session in Spain's schools.

The pre-test and the post-test both assess bullying and cyberbullying awareness before and after playing *Conectado*. The set of questions included in both tests derive from multiple formal and widely-accepted questionnaires that have been demonstrated effective in the school population of Spain (Álvarez García et al., 2013; Garaigordobil & Aliri, 2013; Ortega-Ruiz et al., 2016). In total, the pre-test and the post-test included 18 7-point Likert questions, eliciting how much players agree with each of 18 statements on bullying and cyberbullying. The questionnaire has a Cronbach's alpha of 0.95. The score of each test is calculated as the mean of all answers; therefore, possible test scores range from 1 to 7. Questionnaires and interaction data were managed with *Simva* (Alonso-Fernández, Pérez-Colado, et al., 2019; Pérez-Colato et al., 2019), a tool to simplify experiments to validate serious games and/or assess their players. The game sends questionnaires responses as well as in-game interaction data to *Simva* for their storage. Researchers can then access all collected data, conveniently linked by pseudonymous identifiers assigned to each participant, for further analysis.

As well as the responses to both questionnaires, in-game interaction data (traces) were collected during the experiments, including for example interactions with game characters and objects, and general progress within the game's fictional first week at school. All traces were represented using the xAPI-SG Profile. A tracker component embedded into the game prepared these traces as xAPI-SG statements, sending them to an external server, which, in our case, was integrated with *Simva*.

The Experience API (xAPI) (ADL Initiative, 2016) is a format to capture data from e-learning environments based on activity streams (Snell et al., 2011), and was created by a community lead by ADL. Each xAPI trace, also called a *statement*, represents an in-game interaction. Statements in xAPI are

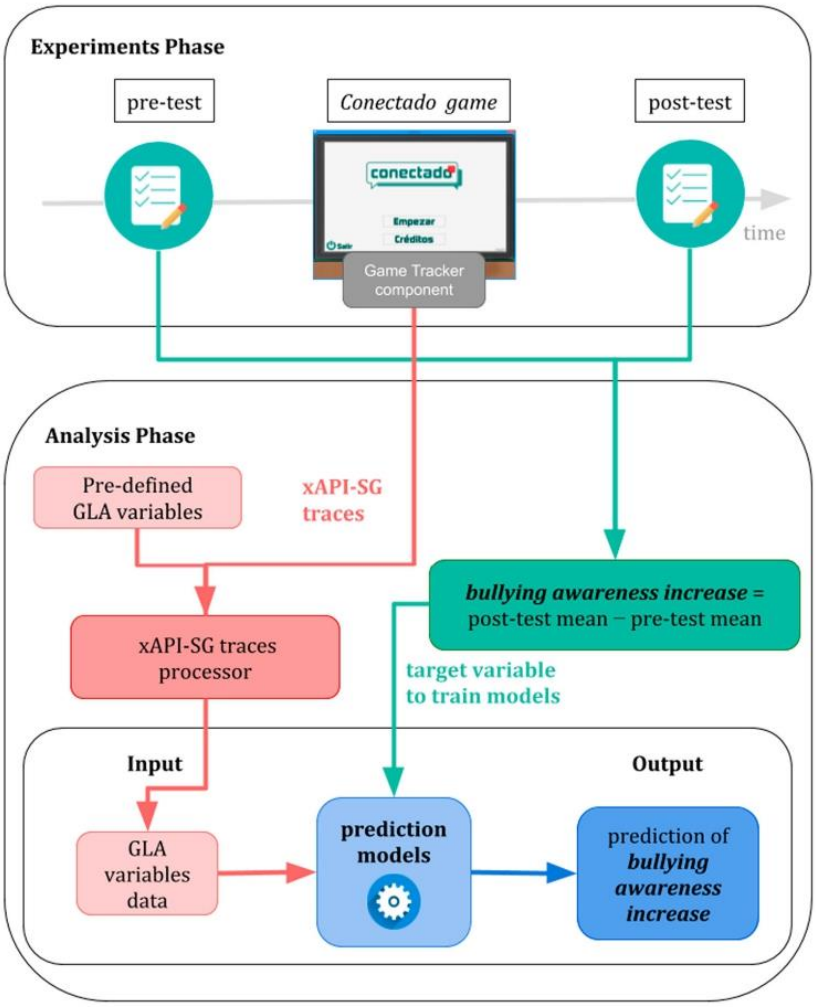

Figure 1. Methodology: (1) During the experiments phase, pre-post questionnaires and interaction data from *Conectado* are collected (top); (2) The analysis phase starts with the xAPI-SG traces, from which GLA variables are derived. Then, the GLA variables are used as input for prediction models of the bullying awareness increase.

formatted as JSON and include three main fields – an actor, a verb and an object – and may include additional ones, such as timestamps or results. Domain-specific profiles can be created to use xAPI in specific contexts and/or communities of practice. For serious games, the xAPI-SG Profile (Serrano-Laguna et al., 2017) was created by the e-UCM Research Group in collaboration with ADL, and includes specific verbs that refer to common structural and design elements found in serious games.

The xAPI-SG traces collected from *Conectado* include, as actor, a pseudonymous identifier provided to each player. The verbs used describe relevant in-game interactions: for instance, "initialized" and "completed" are used, respectively, to track the start and end of in-game days. Figure 2 shows an example xAPI-SG trace from *Conectado*, representing an interaction of the player whose identifier

```
{
  "actor": {
    "name": "user-identifier"
  },
  "verb": {
    "id": "http://adlnet.gov/expapi/verbs/interacted"
  },
  "object": {
    "id": "url-of-game-website/game-version/ComputerOnDesk",
    "definition": {
      "type":
      "https://w3id.org/xapi/seriousgames/activity-types/game-object",
    }
  },
  "result": {
    "extensions": {
      "GameDay": 1.0,
      "GameHour": "21:30",
      "MobileMessages": "True"
    }
  },
  "timestamp": "2019-05-17T12:04:56.835Z"
}
```


Figure 2. xAPI-SG example statement (trace) tracked from an interaction in *Conectado* when the player uses the in-game computer. Text in italics would be replaced by actual identifiers.

appears in the "actor" field, with an in-game object identified as "ComputerOnDesk" (a computer in the game that the player can interact with). The timestamp of the action is included, and the results contain additional information such as the in-game day that the trace belongs to.

The xAPI-SG statements are then processed to derive the GLA variables to be used in the analysis. For each type of statement, we store the following information:

- For "accessed" actions, an identifier for the target, such as "school_bathroom"
- For "initialized" actions, an identifier for the object of the action, such as the full game or a specific in-game day; and a timestamp.
- For "completed" actions, similar information to that of "initialized"; and, if the full game has been finished, we also store the specific ending reached within the result field.
- For "interacted" actions, the target, which may be an in-game element, for example when using items; or, as is the case for conversations, the character that the player conversed with.
- For "progressed" actions, an object identifier. For example, when tracking the changes in variables that represent the level of friendship with other characters, the identifiers of those characters are used.
- For "selected" actions, the object and the results of the action to track in-game decisions. For example, when players can choose to mention the ongoing bullying to parents, the results would include the player's choice, and the object would identify the point where that choice was taken.

With this information, we can calculate the values of the GLA variables described in the following section.

Identifying game learning analytics variables

Once that all the relevant in-game interactions have been collected as xAPI traces, they can be processed to extract GLA variables in a process known as feature extraction. Although this process is not

straightforward, and to a significant extent relies on having in-depth knowledge of the serious game and the exact traces that it generates, we can provide several guidelines and recommendations based on our experience and previous work regarding use of GLA (Alonso-Fernández, Cano, et al., 2019). First, if available, the feature extraction process should be based on the games' Learning Analytics Model (LAM) (Pérez-Colado et al., 2018). The LAM provides the relevant information (events) to be tracked from the game, relating that data with the specifications of the (educational) game design, and describing how that data should be analyzed and/or interpreted. This process could be also guided by a learning analytics reference model (Chatti et al., 2012). If a LAM is not available, we recommend the use of default analysis and visualizations (Alonso-Fernández et al., 2017) based on the main fields of the xAPI-SG Profile, which expose the game's structure and can quickly identify important features to be considered for analysis. Another option is to consult the game designers, who may suggest key elements to analyze from the game.

In *Conectado*, the possible interactions include making choices in the dialogs with characters, such as confronting the bully (Alejandro), mocking a previously bullied character (Maria), or sharing your phone password with bystanders who have also started to bully you. The player can also choose to mention ongoing bullying with the main character's parents or the teacher. After each of these interactions, the level of friendship with other in-game characters can increase or decrease, acting as an indicator of the player's standing regarding those characters. The player can also move within the game's scenarios and interact with objects; for instance, entering the school bathroom when asked to do so by a classmate, attempting to remove chewing-gum placed on clothes by a bully, or using different simulated devices and their (cyberbullying relevant) social applications, such as instant-messaging or social-network apps.

Based on *Conectado*'s LAM, we extracted a set of relevant GLA variables for this game. Some of these variables, such as durations or player-entity interactions, could be obtained directly from xAPI-SG Profile statements, without need for further game-specific knowledge. We additionally included variables representing the in-game actions related to bullying and cyberbullying. From the set of xAPI-SG traces collected from students' gameplays, we extracted the information to fill these pre-defined variables to later use them as input for the prediction models. As an example, variables with the duration of each in-game day for a player are obtained by subtracting the timestamps of the pair of xAPI-SG traces that mention each day as being "initialized" and "completed". Binary variables are extracted by looking up traces that mention their corresponding object identifiers, and then using the results of those traces to arrive at a true or false value. Values of discrete variables are progressively increased as "interaction" traces with the corresponding targets appear.

The 44 GLA variables that we chose to include in our analysis (derived from the xAPI-SG statements) are described in Table 1, including their names, types and brief descriptions. Related variables are described together in the same row of Table 1; for instance, the 5 variables containing the duration of each game day are described together as "duration_day_d, *d* in *[1,2,3,4,5]*". The resulting GLA variables were then used as input for the prediction models described in the next section.

Prediction models of bullying awareness increase

Our final goal is to use the gameplay traces to predict the increase in bullying awareness as a result of playing the game. We define the bullying awareness increase as the difference between the post-test mean score and the pre-test mean score for each player. Therefore, this continuous variable is the target variable for prediction models.

We have used different prediction models to predict the exact value of the increase in bullying awareness and compared predicted results with those obtained in the pre-post-tests. As prediction models, we chose: linear regression, regression trees, Bayesian regression, Support Vector Machines for Regression (SVR), k-nearest neighbors (k-NN), neural networks, random forests, AdaBoost, and gradient boosting. All models were tested with 10-fold cross validation. For all models, different parameters were tuned to find the best ones.

Table 1. Variables derived from the xAPI -SG statements of *Conectado*.

Variable name	Type	Description
accepted_c, *c* in *[Alison, Guillermo, Jose]*	true/false	Player has accepted a friendship request on in-game computer of character *c*
accessed_bathroom	true/false	Player has accessed the school bathroom
confront_Alejandro	true/false	Player has confronted Alejandro
duration	continuous	Total time playing *Conectado* (in minutes)
duration_day_d, *d* in *[1,2,3,4,5]*	continuous	Total time playing day *d* of *Conectado* (in minutes)
ending_number	categorical	Ending reached by the player: 1 for worst ending, 2 for regular, and 3 for best ending
find_earring	true/false	Player has helped Alison to find her earring
friendship_decrease_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]*	discrete	Number of times the player has decreased the level of friendship with character *c*
friendship_increase_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]*	discrete	Number of times the player has increased the level of friendship with character *c*
gum_washed	true/false	Player has washed the gum from the clothes
has_ended_game	true/false	Player has ended the full *Conectado* game
interactions_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Mother, Father]*	discrete	Number of interactions the player has carried out with character *c*
mock_Maria	true/false	Player has mocked Maria
shared_password	true/false	Player has shared the password with classmates
tattle_to_parents	true/false	Player has mentioned bullying to parents at home
tattle_to_teacher	true/false	Player has mentioned bullying to teacher at the school
used_computer	true/false	Player has used the computer at home
used_friends_app	true/false	Player has used social network app on smartphone
used_mobile_chat	true/false	Player has used instant messaging on the smartphone

Results

For each of the 9 prediction models, Table 2 shows the mean absolute error (MAE) and the standard deviation (SD) (normalized to scale [0–10]) for the predictions with the best combination of parameters found for that model.

The model that provides the best results is a Bayesian regression, closely followed by a gradient boosting model, with random forests and Adaboost models at very similar error levels, and all other models providing acceptable results. The difference between the best models is not significant. For the remainder of this section, we focus on the variables that have proven to be most relevant in the best-performing models, which we have identified and attempted to relate to the game design to explain why they appear to have such a great influence on changes in bullying awareness:

- Number of interactions with the character Jose (*interactions_Jose)*: a higher number of interactions predicts higher bullying awareness increase. We consider that a high number of interactions with any character may be a result of a high immersion of the player in the game. This may therefore result in a higher increase in the awareness of the problem.

Table 2. Results of predictions of bullying awareness increase.

Prediction model	Mean Absolute Error (MAE) *normalized to scale [0–10]*	Standard Deviation (SD) *normalized to scale [0–10]*
Linear regression	0,581	0,047
Regression trees	0,557	0,055
Bayesian ridge regression	0,540	0,053
SVR	0,556	0,051
kNN	0,578	0,048
Neural Networks	0,557	0,050
Random Forests	0,551	0,052
Adaboost	0,551	0,057
Gradient boosting	0,548	0,052

- Ending reached (*ending_number)*: a better ending in the game (higher value of ending number) predicts higher increase in awareness. The ending reached is the result of the in-game actions and decisions taken, therefore, it relates to players' behavior in the game. An adequate behavior in a bullying and cyberbullying situation shows a higher awareness of players, either from previous training or from the gameplay, which may show a higher inclination to be attentive in the game and therefore further increase their awareness.
- Duration of in-game day 4 (*duration_day_4*): a higher duration predicts higher increase in awareness. We hypothesize that the specific content of that day may be more relevant for the increase in awareness. Revising its content, it includes the threat to change the player's password, a stolen smartphone, and a case of identity theft. These issues may be more important for players in the target group (12–17 years old), and spending more time on this day, and therefore experiencing them in greater detail, may have a greater impact when increasing awareness.
- Duration of in-game day 3 (*duration_day_3*): in contrast to the previous, duration in day 3 predicts lower awareness increase. We consider that above-average durations may show players losing attention in the game. Revising its content, we notice that the social media has a higher presence, as the main character sees an edited picture of him/herself with classmates making fun of it on the comments. The overlong time on this day may reflect players being distracted by the in-game social media application (which is used as a scripted story element, with minimal actual functionality); and trying to carry out more actions in this app, such as replying to comments may have detracted from gameplay as a whole.

It is important to notice that the information on the prediction relevance of variables, and therefore their discussion and explanation, has only been possible because some of the chosen prediction models allowed the relevance of the input variables to be queried. This would not have been possible for other black-box models, which raises the widely-discussed need of explainable artificial intelligence (xAI), as discussed in (Adadi & Berrada, 2018; Carvalho et al., 2019).

Discussion

This case-study has showcased the process of performing an evidence-based evaluation of the effect of a serious game, based on in-game user interaction data. The use of the xAPI-SG Profile standard to collect traces proved to be very useful to simplify the evaluation, as it not only simplified the process of collecting and storing traces, but also that of processing and analyzing data. The xAPI-SG is a standard yet powerful and simple format that allowed easy extraction of the information that we were interested in.

We consider that the approach used with *Conectado* can be generalized to other serious games or, at least, to other linear, narrative serious games. The case study builds upon previous work (Alonso-Fernández et al., 2020), in which, using a serious game to teach first aid techniques, the goal was to predict players' final knowledge about the topics covered in the game. In this experience, interaction data was collected and analyzed to derive relevant variables containing information to use as input to predict players' knowledge.

From the common points encountered on the current and previous experience, we consider that this process could be generalized to carry out other evidence-based evaluations of the effectiveness of serious games. The steps followed can be generalized, first, using a standard to track in-game interactions such as the xAPI-SG Profile. Once interaction data are collected, a further step towards generalization is to gather an initial set of variables to derive from the xAPI-SG traces, based on available fields such as the duration of in-game activities, and interactions with relevant in-game items and characters; which can later be complemented with game-dependent information. The initial set of variables can be used as a baseline of what game learning analytics can conclude for the serious game and can be extracted automatically if analytics traces are formatted using an xAPI-SG Profile

representation. With those GLA variables, we recommend testing interpretable prediction models that provide information of the relevance of each variable, such as tree-based models (random forests, gradient boosting), which can help to interpret and inform the evaluation process and its results. Moreover, using xAPI allow SGs' developers and researchers to build and reuse a tooling ecosystem for both statements gathering, analysis and predictions.

There are some limitations to the generalizability of this study. First, the fact that the videogame has a narrative, almost-linear structure and a low playing time restricts the variability of the interactions for players. Second, the discussion of the relevance of specific variables in our results is limited by the fact that the prediction model is not a black-box model. Finally, selection of GLA variables is not straightforward and could limit the generalization of our approach; however, we have provided guidelines and recommendations to identify an initial set of GLA variables for use in prediction.

Conclusions

We have showcased a full example of an evidence-based evaluation of the effectiveness of a serious game from interaction data: from the possible interactions in the game, their collection using the xAPI-SG standard, the analysis of the resulting traces to derive variables, and the use of those variables as input for prediction models, which, once trained (with pre-post experiment data), provide evaluations of increase of bullying awareness in players based solely on in-game interaction data.

The experiments phase has shown the convenience and advantages of using a standard data format to simplify collection, processing and analysis of interaction data. As the xAPI-SG format is well-defined, no further processing is required, and its use greatly simplifies the feature extraction process. The analysis phase has described the process of bridging the gap between in-game interactions (collected as xAPI-SG traces) and relevant information (stored as GLA variables). The feature extraction was performed based on our knowledge from the game, although many variables could have been selected based solely on the fields available in the xAPI-SG Profile, since the use of this standard format also exposes the general structure of games where it is used. The explainability of the results, which in our case was possible due to our choice of prediction models, is especially relevant towards the deployment of our approach in educational scenarios, as it is required to be able to explain how the results have been obtained, both to teachers and to students evaluated with such techniques.

It is also important to notice that the serious game used on this case-study was not originally designed to be an assessment tool; its mechanics and interactions were solely designed to depict a bullying situation which would make players reflect on their actions and their consequences. Some changes in the game could improve the results: a higher variety of options, leading to even more endings, could provide deeper insight on the actions that players will take in similar situations, at the expense of an increase in game complexity. In the same sense, a broader set of interactions with the characters could provide more information about players' attitudes with the different profiles (e.g. bully, observer, person previously bullied), although this would again increase the game's complexity.

Based on the current and previous work, we consider that most of the steps that we followed could be further generalized, using the xAPI-SG standard to capture interaction data and defining a minimum set of variables to extract from xAPI-SG traces, thus establishing a standard, automated process for evidence-based evaluations of the effectiveness of serious games.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been partially funded by Regional Government of Madrid (eMadrid S2018/TCS-4307, co-funded by the European Structural Funds FSE and FEDER), by the Ministry of Education (TIN2017-89238-R), by the European Commission (Erasmus+ IMPRESS 2017-1-NL01-KA203-035259) and by the Telefónica-Complutense Chair on Digital Education and Serious Games.

Notes on contributors

Cristina Alonso-Fernández obtained her Bachelor in Computer Science and her Bachelor in Mathematics for the Complutense University of Madrid in 2016. A year later, she finished the Master in Data Mining and Business Intelligence, also for the UCM. Since September 2016, she is part of eUCM, where she has worked for the H2020 Beaconing Project. She is currently doing her PhD in Computer Science. Her research interests include educational videogames and application of data analysis and data mining for their improvement.

Antonio Calvo-Morata obtained his bachelor in Computer Science for the Complutense University of Madrid in 2014. In 2017, he completed the Master in Computer Science, also in the Complutense University. He is currently doing his PhD in Computer Science. Antonio has been part of the research group e-UCM since 2014, as a contract researcher for projects eMadrid and H2020 RAGE. His research interests include the study of educational videogames and their application in schools, as well as the use of Learning Analytics techniques to improve their efficacy and their validation as an educational tool.

Manuel Freire has a PhD in Computer Science from the Universidad Autónoma de Madrid (UAM). He is interested in Information Visualization, Human-Computer Interaction, Online Learning, Serious Games, and Plagiarism Detection. He performed a 2008 post-graduate Fulbright scholarship in the Human-Computer Interaction Lab (HCIL-UMD), working with Ben Shneiderman and Catherine Plaisant. In 2010, he became a member of the e-UCM group in the Universidad Complutense de Madrid (UCM), where since 2013 he is an Associate Professor.

Iván Martínez-Ortiz works as Associate Professor in the Department of Software Engineering and Artificial Intelligence (DSIA) at the Complutense University of Madrid (UCM). He has been assistant to the Vice-Rector of Technology at UCM and Vice-Dean for Innovation in the Computer Science Studies. He has been Lecturer in the Computer Science School at UCM in the Computer Science School at the Centro de Estudios Superiores Felipe II. He received a Bachelor in Computer Science (first in the Dean's List ''premio extraordinario'') and a Master and PhD in Computer Science from the UCM. His research interests include e-learning technologies and the integration of educational modeling languages, serious games and e-learning standardization.

Baltasar Fernández-Manjón received the PhD degree in physics from the Universidad Complutense de Madrid in 1996. He is a Full Professor of computer science at UCM and director of the e-UCM e-learning research group. He has the Honorary Complutense-Telefonica Chair on Digital Education and Serious Games. His research interest is focused on the applications of ICT in education and in serious games and educational simulations applied to different domains (e.g. medicine, education). He is also working in game learning analytics and the application of e-learning standards to the integration of those technologies in e-learning systems. He is a senior member of the IEEE.

ORCID

Cristina Alonso-Fernández http://orcid.org/0000-0003-2965-3104
Antonio Calvo-Morata http://orcid.org/0000-0001-8701-7582
Manuel Freire http://orcid.org/0000-0003-4596-3823
Iván Martínez-Ortiz http://orcid.org/0000-0001-6595-5690
Baltasar Fernández-Manjón http://orcid.org/0000-0002-8200-6216

References

Adadi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

ADL Initiative. (2016). *xAPI specification*. 2014. https://github.com/adlnet/xAPI-Spec/blob/a752217060b83a2e15dfab69f8c257cd86a888e6/xAPI.md

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612

Alonso-Fernandez, C., Calvo, A., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2017, April 25–28). *Systematizing game learning analytics for serious games*. 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, 1111–1118. IEEE. https://doi.org/10.1109/EDUCON.2017.7942988

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301–309. https://doi.org/10.1016/j.chb.2019.05.036

Alonso-Fernández, C., Pérez-Colado, I. J., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). *Using* Simva *to evaluate serious games and collect game learning analytics data*. LASI Spain 2019: Learning Analytics in Higher Education (pp. 22–34). http://ceur-ws.org/Vol-2415/paper03.pdf.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, *36*(3), 350–358. https://doi.org/10.1111/jcal.12405

Álvarez García, D., Núñez Pérez, J., & Dobarro, A. (2013). Cuestionarios para evaluar la violencia escolar en Educación Primaria y en Educación Secundaria: CUVE3-EP y CUVE3-ESO. *Apuntes de Psicología*, *31*(2), 191–202.

Calderón, A., & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, *87*, 396–422. https://doi.org/10.1016/j.compedu.2015.07.011

Calvo-Morata, A., Freire-Moran, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2019). Applicability of a cyberbullying videogame as a teacher tool: Comparing teachers and educational sciences students. *IEEE Access*, *7*, 55841–55850. https://doi.org/10.1109/ACCESS.2019.2913573

Calvo-Morata, Antonio, Rotaru, Dan Cristian, Alonso-Fernandez, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernandez-Manjon, B.. (2020). Validation of a cyberbullying serious game using game analytics. *IEEE Transactions on Learning Technologies*, *13*(1), 186–197. https://doi.org/10.1109/TLT.2018.2879354

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5/6), 318. https://doi.org/10.1504/IJTEL.2012.051815

Freire, M., Serrano-Laguna, Á, Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology* (pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_21-1

Garaigordobil, M., & Aliri, J. (2013). Ciberacoso (''cyberbullying'') en el País Vasco: Diferencias de sexo en víctimas, agresores y observadores. *Behavioral Psychology/ Psicologia Conductual*, *21*(3), 461–474.

Kato, P. M., & Klerk, S. D. (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology*, *18*, 1–6.

Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning analytics as an assessment tool in serious games: A review of literature. In M. Ma & A. Oikonomou (Eds.), *Serious games and edutainment applications* (pp. 537–563). Springer International Publishing. https://doi.org/10.1007/978-3-319-51645-5_24

Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Evaluar el bullying y el cyberbullying validación española del EBIP-Q y del ECIP-Q. *Psicología Educativa*, *22*(1), 71–79. https://doi.org/10.1016/j.pse.2016.01.004

Peng, W., Lee, M., & Heeter, C. (2010). The effects of a serious game on role-taking and willingness to help. *Journal of Communication*, *60*(4), 723–742. https://doi.org/10.1111/j.1460-2466.2010.01511.x

Perez-Colado, I., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2018, April 17–20). *Game learning analytics is not informagic!*. 2018 IEEE Global Engineering Education Conference (EDUCON), Tenerife, Spain, 1729–1737. IEEE. https://doi.org/10.1109/EDUCON.2018.8363443

Perez-Colado, I. J., Calvo-Morata, A., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2019, July 15–18). *Simva: Simplifying the scientific validation of serious games*. 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Maceió, Brazil, 113–115. IEEE. https://doi.org/10.1109/ICALT.2019.00033

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, *50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

Shoukry, L., Göbel, S., & Steinmetz, R. (2014). *Learning analytics and serious games: Trends and considerations*. Proceedings of the 2014 ACM International Workshop on Serious Games. https://doi.org/10.1145/2656719.2656729

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

Shute, V., & Ventura, M. (2013). Stealth assessment. In J. Michael Spector (Ed.), *The SAGE encyclopedia of educational technology* (p. 91). SAGE Publications. https://doi.org/10.4135/9781483346397.n278

Snell, J., Atkins, M., Norris, W., Messina, C., Wilkinson, M., & Dolin, R. (2011). JSON activity streams 1.0. *Search ResultsWeb resultsActivity Streams Work*, *22*(8), 2013.

Xu, Y., Johnson, P. M., Lee, G. E., Moore, C. A., & Brewer, R. S. (2014). Makahiki: An open source serious game framework for sustainability education and conservation. *International Association for Development of the Information Society*, *8*, 131–138.

6.1.4. Improving evidence-based assessment of players using serious games

Full citation

Cristina Alonso-Fernández, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2021): **Improving evidence-based assessment of players using serious games.** Telematics and Informatics. (in press). DOI: 10.1016/j.tele.2021.101583.

Impact metrics: JCR 2019, Impact Factor: 4.139, Q1 in Information Science & Library Science.

Abstract

Serious games are highly interactive systems which can therefore capture large amounts of player interaction data. This data can be analyzed to provide a deep insight into the effect of the game on its players. However, traditional techniques to assess players of serious games make little use of interaction data, relying instead on costly external questionnaires. We propose an evidence-based process to improve the assessment of players by using their interaction data. The process first combines player interaction data and traditional questionnaires to derive and refine game learning analytics variables, which can then be used to predict the effects of the game on its players. Once the game is validated, and suitable prediction models have been built, the prediction models can be used in large-scale deployments to assess players solely based on their interactions, without the need for external questionnaires. We briefly describe two case studies where this combination of traditional questionnaires and data mining techniques has been successfully applied. The evidence-based assessment process proposed radically simplifies the deployment and application of serious games in real class settings.

Full publication

Telematics and Informatics 60 (2021) 101583

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Improving evidence-based assessment of players using serious games

Cristina Alonso-Fernández *, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón

Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, C/ Profesor José García Santesmases, 9, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Data science applications in education
Evaluation methodologies
Games
Teaching/learning strategies

ABSTRACT

Serious games are highly interactive systems which can therefore capture large amounts of player interaction data. This data can be analyzed to provide a deep insight into the effect of the game on its players. However, traditional techniques to assess players of serious games make little use of interaction data, relying instead on costly external questionnaires. We propose an evidence-based process to improve the assessment of players by using their interaction data. The process first combines player interaction data and traditional questionnaires to derive and refine game learning analytics variables, which can then be used to predict the effects of the game on its players. Once the game is validated, and suitable prediction models have been built, the prediction models can be used in large-scale deployments to assess players solely based on their interactions, without the need for external questionnaires. We briefly describe two case studies where this combination of traditional questionnaires and data mining techniques has been successfully applied. The evidence-based assessment process proposed radically simplifies the deployment and application of serious games in real class settings.

1. Introduction

Serious Games (SGs) are games that "do not have entertainment, enjoyment or fun as their primary purpose" (Michael and Chen, 2005). Digital serious games provide an engaging, highly interactive environment with many possibilities for causing an effect on players (Dörner et al., 2016). SGs also present an opportunity for proactive learning by involving players/learners in an immersive learning experience where they can apply their knowledge, learn from experience, and test complex or risky scenarios in a safe environment. Due to these features, SGs have been successfully applied in varied domains such as medicine, the military, or complex processes training, among others; example success-cases include dealing with phobias (Donker et al., 2018), medical training (Boada et al., 2016), and in-company training (Michael and Chen, 2005).

In SGs, a wide and diverse range of interactions can be tracked and analyzed to gain insight into their players' behaviors. Collection of interaction data is widespread in the games industry. It forms the basis of the field of Game analytics (GA), defined as the application of analytics for "game development and game research" aiming to provide "support for decision-making at all levels (…) from design to art, programming to marketing, management to user research" (El-Nasr et al., 2013). Data collection has also been applied in education

* Corresponding author.
E-mail address: crisal03@ucm.es (C. Alonso-Fernández).

https://doi.org/10.1016/j.tele.2021.101583
Received 9 December 2020; Received in revised form 21 January 2021; Accepted 4 February 2021
Available online 13 February 2021

via Learning Analytics (LA), "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long et al., 2011), mainly focusing on predicting students' success and providing feedback (Gašević et al., 2015). The combination of GA and LA techniques in the context of serious games is called Game Learning Analytics (GLA) (Freire et al., 2016). GLA can help better understand the player game experience and characteristics, allowing the game design to be adapted and improved based on this player interaction evidence.

Assessment of players in serious games is usually performed with formal external measures, typically gathered via questionnaires, while Game Learning Analytics (GLA) data is rarely included in the assessment process (Alonso-Fernández et al., 2019). From our experiences evaluating SGs and assessing players using them, we propose a new approach to assess players using serious games based on GLA data. Once a game is validated using traditional external questionnaires, players' interactions in the game are analyzed to obtain representative GLA variables and prediction models that can be used to predict serious games' effect on players, as measured by the external questionnaires. For further deployments, the chosen prediction models can automatically assess players in a non-intrusive way, solely based on their interactions. This type of assessment becomes particularly useful in large-scale game applications in real domains, as it avoids potentially costly external assessment of players. The information gathered can also be used to improve and adapt the game to players' characteristics and provide targeted and personalized feedback.

We propose an assessment of players using serious games based on their game interactions: the evidences collected from their gameplays provide the game learning analytics data that can be used to reliably predict the effect of the game on players.

The rest of this paper is structured as follows: Section 2 reviews related work on serious games evaluation and the assessment of their players; Section 3 describes our evidence-based process for assessment of players using serious games based on the collection and analysis of game learning analytics data from their interactions to predict the game's effects; Section 4 briefly presents two case studies, with different serious games, where we have tested the previous process; Section 5 discusses our process and its limitations; finally, Section 6 summarizes the conclusions of our work.

2. Related work

Serious games are commonly validated with questionnaires (Calderón & Ruiz, 2015) where players are asked to complete formal external questionnaires both before (pre) and after (post) the gameplay. The results from both questionnaires are then compared, and the game is considered to be effective if the difference between their results is statistically significant. This widely accepted methodology has some drawbacks: first, the questionnaires must be previously validated to ensure that they provide a reliable measure of the characteristics the game aims to change. Second, use of questionnaires significantly increases the effort and time in preparation, administration, and analysis for teachers or researchers who apply them. As a final drawback, the assessment is performed outside the game environment, breaking player immersion and requiring additional mechanisms to deliver the questionnaire, collect player responses, and link them back to their authors.

A data-based, or at least data-informed, evaluation approach, taking advantage of the potential of game learning analytics data collected from players' interactions, could provide more authentic and precise evaluation metrics. These metrics could be analyzed for deeper insight, both while the game is being played (in real-time), or after the gameplay is complete. Analyzing the data, patterns can be discovered, both from single-player interactions and by combining data from multiple users (Shoukry et al., 2014). For instance, player profiles could be created based on their game preferences to help SG designers tailor their games or model players' exploration to understand their learning pathways better. Visual analytics can also help to gain insight into players' behavior (Wallner & Kriglstein, 2015) and provide teachers with real-time information via learning analytics dashboards, for instance, to assist players as they play (Charleer et al., 2018).

Evidence-based approaches are also being used to assess players, as different studies have started to investigate how their performance can be assessed directly from user-generated data, instead of relying on external questionnaires (Loh & Sheng, 2015). The field of *stealth assessment* (Shute et al., 2017) aims to embed the assessment in a non-intrusive manner into the gaming environment. Assessment is then based on what players do in the game, as opposed to the use of immersion-breaking external questionnaires. This approach is based on the collection of specific data from players' gameplays, which are stored in log files. These collected high-level metrics are then analyzed and correlated with players' knowledge (Shute et al., 2013). The use of games for assessment has consequently drawn the attention of many recent research studies. In fact, the most common application of data science to game learning analytics data is that of assessing students, either to measure learning or to predict performance (Alonso-Fernández et al., 2019). The work of (Halverson & Owen, 2014) presents a model for game-based assessment that collects data from keystrokes and clicks, and identifies 15 moments in a game as evidences of learning to correlate with learning gains. The model is exemplified with a science game, for which results showed that the type of mistakes made were the best learning predictors, even more so than the number of times that players played, or the number of successes or failures experienced. Other studies propose the use of higher-level metrics obtained from in-game interactions such as learning observables (Serrano-Laguna et al., 2017) or variables with aggregated learning analytics data (Alonso-Fernández et al., 2019). Research has also been conducted on game-based assessment of specific 21st century skills, such as persistence (Dicerbo, 2013). From log data, researchers can identify players with specific goals and then create measures of persistence toward those goals, including information such as progress and time spent completing difficult tasks. Game-based assessments are being incorporated in other related fields. For instance, gamified applications for language learning are including machine learning to create computer-adaptive assessments (Settles and Laflair, 2020).

Despite these studies, research conducted on games as tools for assessment is still limited (Homer et al., 2018). While few serious games for learning have been primarily built for assessment (Sliney & Murphy, 2011), several authors consider it important to include assessment as part of the design phase (Ifenthaler et al., 2012). Another critical issue is that, so far, studies have focused on game-based

assessments on a case-by-case basis. This results from each serious game having different goals, structure, and contents, therefore providing different opportunities for assessment. For each game, or type of game, different kinds of interaction data will be available for collection, at different levels of granularity, and offering different evidences for assessment. Therefore, research is needed to shed light on how game features and categories contribute or detract to their validity from the point of view of assessment (Kato & Klerk, 2017). These features and characteristics that could relate to players' assessment will be tightly related to the game design and learning design. In the literature review of (Liu et al., 2017), authors found out that serious game features and metrics were primarily used for learner performance and game design strategies, and highlighted the need for more data-based research studies on this topic. Authors have carried out several studies trying to leverage the costly process of creating game-based assessments. The work of (Kim et al., 2019) proposes a process for game designers and developers to create games for educational assessment, balancing assessment needs and the gameplay experience through the phases of design, development and evaluation. This process comprises the collection of evidences from the game, the analysis of such evidences to create relevant variables, and the evaluation of the model.

The generalization of evidence-based players' assessments playing serious games is key towards their widespread use and future impact. For this, better data collection and sharing are vital parts of a continuous process to improve teaching and learning (Shute & Rahimi, 2017). For this process to be generalized, it seems essential to combine data collection and analysis with standard and systematic processes. Simultaneously, serious game designers, developers, and researchers need access to tools that can capture educationally relevant data from their games, and even more importantly, tools that can analyze collected data to yield insights into the progress and actions of players in those games. This information can then also be used for players' assessment.

We propose an evidence-based process to assess players using serious games, described in detail in the following section. Taking advantage of well-known data science techniques, our approach uses data collected during game validation to create models that can predict the effect of the game on its players. Use of this approach greatly simplifies game-based assessments, which are currently limited and conducted on a case-by-case basis, while retaining the advantages of evidence-based assessment. We have only tested this approach with narrative adventure videogames, but we consider the steps to be generalizable to other similar genres, such as geo-localized videogames with a narrative component.

3. Evidence-based assessment of players using serious games

Our proposed evidence-based assessment process combines in-game player interaction data with the traditional external data collected from questionnaires during formal game validations. Fig. 1 provides an overview of our proposed evidence-based assessment process. It comprises the following phases and tasks:

1. Game validation phase.
   a. Collect common player interactions using a standard and validated format (Section 3.1).
   b. Analyze the collected traces to choose an initial set of variables containing GLA information based on the Learning Analytics Model, and/or game designers' guidance, and refined through exploratory analysis and visualizations in our data science environment called T-Mon (see *feature extraction* process, Section 3.2).
   a. Use the selected variables as input of the prediction models to measure the impact of games on their players, and the pre-post questionnaires as target output variable to be predicted by the models. Predictions are used to validate the models, and the predictive relevance of the variables used provides feedback to iterate the process, if needed (Section 3.3).

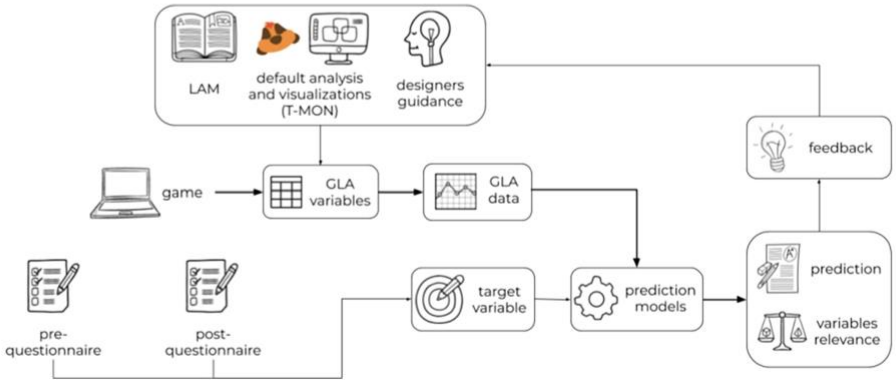

Fig. 1. Full process for evidence-based assessment of serious game players (tasks during game validation phase): the game interaction traces collected fill the pre-defined set of GLA variables to be used as input for the prediction models. The target variable used for prediction is based on pre-post results.

2. Game deployment phase.
   a. After the validation, large-scale deployments can be conducted where players are assessed based solely on their game interactions (Section 3.4).

3.1. Collecting players data: pre-post questionnaires and in-game interaction data

In our process, the first step is to collect the data to assess players with. For this purpose, we need to collect both pre-post questionnaires (or any other validated measure to be used as the target value for the predictions) as well as interaction data. Questionnaires should be formally validated by experts in the domain, to ensure that they provide a reliable measure of the characteristic that the game seeks to affect, such as awareness or knowledge.

Although the type of data that can be collected from a serious game will depend on its content, structure and features, there are some common interactions in game analytics (GA) and learning analytics (LA) that can be extrapolated to serious games. GA data relates to game design and structure: number of clicks, avatar location in the game environment and characteristics, movements and changes of scenes or levels, items used, total time spent in the game, interactions with interface elements and non-player characters, points scored, in-game selections, and quest completions. These data can also be combined to derive "game metrics". As an example, by combining a "total time" and "points scored" for a given player, we can derive a "points per minute" metric. GLA data relates to the learning design of the games, reflecting information about the learning progress and process of players/learners. Previous work on LA has identified multiple variables with strong predictive power, such as the interactions with elements of the learning environment (e.g. videos or activities) (Brinton et al., 2016; Ruipérez-Valiente et al., 2017). When it comes to games, learning analytics data focuses on in-game player actions that affect learning.

To systematize the specific data to be collected from serious games, we propose the use of the validated and standardized *Experience API for Serious Games Profile* (xAPI-SG) (Serrano-Laguna et al., 2017). The Experience API describes an Application Programming Interface that allows e-learning content, such as serious games, to interact with a Learning Record Store (LRS) that stores the interaction data generated by the learning content. The xAPI-SG profile defines a common shared vocabulary and semantics for Serious Games. Software tools that support it need not be tied to specific SGs, thus allowing their users to reason on and compare data from multiple SG. Some of the commonly used interactions included in the xAPI-SG Profile are: verbs *initialized*, *progressed* and *completed* to measure completion and progress in so-called *completables* of types *serious-game*, *level* or *quest*; verbs *accessed* and *skipped* to collect changes in scenes of types *screen* or *area*; verbs *selected* and *unlocked* to track user-choices when confronted with a *question*, *menu* or *dialog-tree*; and *interacted* and *used* verbs to track interactions with *items* or *non-player-characters*.

The use of a standard data format such as xAPI-SG is a clear benefit when systematizing the collection of traces and their analysis to derive relevant information from user gameplays. Such formats facilitate the integration of tools from different providers and help to comply with personal data-protection laws: art. 20 of the EU GDPR requires data controllers to use a "structured, commonly used and machine-readable format" when users request access to their data, or transfer to other data controllers (European Commission, 2018). Additionally, while standard data collection formats are not commonly reported on the literature (Alonso-Fernández et al., 2019), their uptake would greatly assist in result replication and data sharing. Having a common interchange format also fosters the creation of a tool ecosystem created by different actors.

3.2. Extracting GLA variables from interaction data

Once the raw traces with user interaction data are collected, they can be analyzed to extract higher-level meaningful information about the actions of players within the game. Our process synthesizes the information available in the data traces (collected in xAPI-SG format) into a smaller set of GLA variables. Ideally, the definition of such variables should be described in the game's Learning Analytics Model (LAM) (Pérez-Colado et al., 2018), cooperatively created by both educational experts and game designers. LAMs build on the game's learning design and game design, which define the educational goals of the game and how these are reflected on the specific game design choices taken depending on their educational goals. Based on both designs, a LAM determines the data to be collected from the game and how these data are to be analyzed into GLA variables and interpreted to provide meaningful information about the actions of a player in the game. It also may define any posterior visualization, feedback or reporting to do with the analysis results.

If such a LAM is not available, the game designers may suggest what information to obtain from the game and analyze it into GLA variables. Additionally, analysis and visualizations of collected xAPI-SG traces can provide important insights on the data collected and guide the choice of some GLA variables. For this purpose, we have created our data science environment called T-Mon (a trace monitor in xAPI-SG format). T-Mon contains a set of Python Jupyter Notebooks, available as open source at a GitHub repository[1]. T-Mon's notebooks provide a default set of analyses and visualizations that can be applied to any given JSON file containing xAPI-SG traces: overall game progress; choices in alternatives, and if applicable, the fraction of those considered correct and incorrect; progress, scores and times per game activity or subsection; content seen and skipped; and interactions with game items and areas and over time. The interactive interface allows to filter the data and configure the visualizations to gain a more in-depth insight into the data. T-Mon is intended both to provide quick overviews of collected data and to allow in-depth exploratory analysis to refine the choice of GLA

[1] https://github.com/e-ucm/t-mon

variables that will be used in subsequent steps: the Jupyter Notebooks (Project Jupyter, 2020) T-Mon builds upon are a commonly used tool in data science to perform such analyses and provide access to an extensive and actively maintained collection of utilities to manipulate and explore data (Jupyter Team, 2020).

A further advantage of using xAPI-SG to collect data is that, since xAPI-SG is designed to model and report on essential structures and concepts found in serious games, those structures and concepts are likely to yield a right choice of initial GLA variables. Table 1 proposes a non-exhaustive set of pre-defined GLA variables for each player that can be easily derived from any set of traces that follow the xAPI-SG Profile. Such variables include the number of interactions with each in-game object and character (count of *interacted* traces per object), or the duration of each level/game (difference in timestamp of *completed* and *initialized* traces per object of type serious-game or level). We are currently working on extending the xAPI-SG Profile to include more precise definitions of the required fields in each trace (using *statement templates*) and the required sequence of traces (using *patterns*) to clarify the expected traces to extract such GLA variables. While game-specific variables as specified in a LAM or suggested by expert designer knowledge are of course preferable, a set of ready-to-use generic variables can be highly useful to complement game-specific variables, and allows the use of our process even when no LAM or designers are available. They also constitute a good starting point for refinement using T-Mon.

Once the GLA variables have been chosen, they can be used for player modelling, as they can provide a rich insight into players' actions. Once enough data have been collected, richer information may be obtained via a wide variety of algorithms and techniques, for purposes such as assessment or adaptation. Supervised, unsupervised, or reinforcement learning techniques can be applied to the derived variables. In our case, we use supervised techniques (Soni, 2007) to predict the serious game's effect based on the information found in GLA variables derived from user interactions.

3.3. Creating the prediction models with GLA evidences

The next step is to create the prediction models to accurately measure the effect of the game on its players. By default, we define such effect, which will be the target variable for our models, as the improvement between the scores of the pre- and post- questionnaires, caused by playing the serious game. If we were only interested in measuring the final effect on players after playing, the post questionnaire score alone could be used as the target variable. The prediction models use the previously defined GLA evidences, filled with the data captured during the game validation, as input data.

To consider a serious game effective in educational scenarios, it first needs to be validated, ideally using a formal validation process. We use the formal validation step to create the prediction models that will be used in the deployment phase. The formal validation of the serious game is commonly performed with pre-post questionnaires. The comparison of the results between both questionnaires should, ideally, show a statistically significant difference between pre-questionnaire and post- questionnaire. If such difference is significant, we consider that the game is experimentally validated. During these experiments, we also collect relevant game learning analytics data from players' in-game interactions. With both questionnaires and GLA data, we can create the prediction models that will be used for game effect assessment during the deployment phase. The prediction models take as input the GLA data from players' interactions and predict the improvement (difference between pre- and post- questionnaire results). This process is experimental and can be iterated until accurate-enough models are created, by changing and refining the GLA variables according to their relevance as reported by the results of the prediction models. In our experiments, we have found accuracies above 90% to be achievable, and suggest this figure as a workable goal. Once an accurate-enough level is reached, the final prediction model is retained for the next step of deployment, where it will be used for automatic non-intrusive assessment of players.

For the specific prediction models to be tested, an increasingly broad and varied range of options is available. At least in the first

Table 1
Correspondence of xAPI-SG traces (object type, verb and other fields) to derive GLA variables.

xAPI-SG fields			GLA variables	
Object type	Verb	Other fields	Name	Description
Accessible: area, cutscene, screen, zone	Accessed	Object id	*Accessed_id*	Number of times the accessible *id* has been accessed
	Skipped	Object id (cutscene)	*Skipped_id*	Number of times the cutscene *id* has been skipped
Completable: serious-game, level, quest	Initialized	Object id, timestamp	*Duration_id*	Duration of completable *id* (calculated in combination with *completed* trace of same *id*)
	Progressed	Object id, result progress, timestamp	*Progress_id_time*	Progress in completable *id* per timestamp *time*
	Completed	Object id	*Completed_id*	True if completable *id* has been completed
		Object id, timestamp	*Duration_id*	Duration of completable *id* (calculated in combination with *initialized* trace of same *id*)
		Object id, result score	*Score_id*	Score obtained in completable *id*
Alternative: question, dialog-tree, menu	Selected	Object id (question), result success	*Correct_id*	True if question *id* has been successfully answered
		Object id (dialog), result response	*Response_id*	Response selected in dialog *id*
		Object id (menu), result response	*Selection_id*	Option selected in menu *id*
Target: non-player character, enemy, item	Interacted	Object id	*Interactions_id*	Number of interactions with target *id*
	Used	Object id (item)	*Uses_id*	Number of uses of item *id*

iterations, we recommend using interpretable models (Adadi & Berrada, 2018) that provide information about the relevance of the input variables towards the predictions. This will provide feedback about the importance of specific GLA variables (and, therefore, about users' interactions), allowing us to improve the process before moving to the deployment phase. Linear and tree-based prediction models are a simple baseline to start from. More complex models may improve the results: for instance, ensemble methods based on trees, such as random forest or gradient boosting. These complex models could provide more precise results while still giving feedback about how relevant the input variables are towards the prediction results. The models may then be reused and adapted for different contexts. Traces can be re-examined to generate additional GLA variables or change existing ones based on variable relevance as reported by such models.

The creation of prediction models relies on both questionnaires and interaction data. By collecting player interaction data while the serious game is formally validated, prediction models can be trained with the same questionnaires that are used for formal serious game validation. Suppose the results obtained in the traditional formal-validation questionnaires show a significant improvement on players. In that case, the serious game is formally validated – and models can be built immediately, without the need for further experiments. Instead of only proving the game's efficacy in the chosen educational scenario, we have also built a set of prediction models to be used for students' assessment, and identified a subset of user interaction data that is to be collected from the game in order to make assessment predictions.

The number of users to include in this validation phase is not clear, but considering the reported number of users in other data-based research on serious games (Alonso-Fernández et al., 2019), we recommend including at least 100 users. The information gathered during the validation phase can also be used to improve the game, if the data shows behaviors that do not align with the game design or learning design. The game can be adapted based on players' characteristics, for example, by creating player profiles based on their game behaviors, and then specifically adapting the game to each profile. Collected data can also be used to provide more targeted or personalized feedback to help players progress in the game. The resulting game, updated with features and personalization based on the feedback from the previous round, would then be subjected to another validation phase, leading either to further iterations or to a fully validated game.

3.4. Assessing players in large-scale game deployment

Once the serious game has been formally validated, the deployment phase can start, with the game applied in classrooms and other real-world educational settings. To be able to gather information from users' experience and to assess them based on their interactions, this application should include the collection of data from relevant interactions. The deployment process for large-scale scenarios reads as follows:

1. Students access the SG and play the game from beginning to end. We have used anonymous identifiers that allow only teachers to de-anonymize student data to ensure that privacy requirements are met while still linking questionnaire responses to each student's game-interaction data.
2. A tracking component integrated in the game sends the relevant traces generated from player interactions to the analytics tool while students are playing. The user interaction traces should follow a well-defined format (for instance, the xAPI-SG Profile), as required by the analytics tool that will receive it.
3. The analytics tool takes interaction data as input, uses it to fill the pre-defined GLA variables, and uses them as input to the previously created prediction models, to derive prediction outputs for the students' assessment.

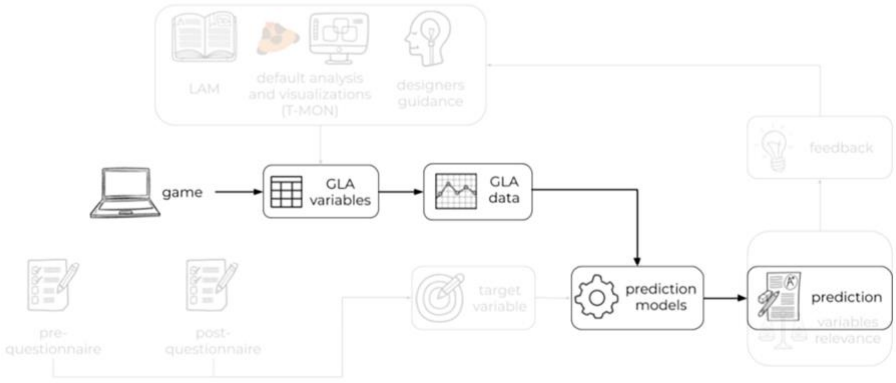

Fig. 2. Full process for evidence-based assessment of serious game players (tasks during game deployment phase): the GLA data derived from game interaction traces is input to the prediction models to obtain a measure to assess players. External questionnaires are no longer required.

4. Once students have finished their gameplay, teachers will receive the predicted score based on each students' in-game interactions, possibly together with another analytics information. They can then use this information, together with any other evaluation of their own, to obtain the final students' assessment.

Note that the prediction models provide the assessment output for students once they have finished playing the game, and therefore once all the input data required by the models is available.

The assessment obtained with this process is therefore automatic and non-intrusive, and simplified from both ends: teacher preparation and execution times typically required for post-game assessment are reduced, and students will simply play a game without the added time, disruption, and pressure of completing the questionnaires. Game-based assessment can also provide institutions and managers means of evaluating the efficacy of games for education and simplify the assessment process of their students. The previously used pre-post questionnaires are no longer required during the large-scale deployment, which simplifies the application of SGs in real-world larger settings. This allows students to play the game for longer periods, and/or teachers to include additional activities related with the gameplay (e.g. discussion, post-game questions), instead of the traditional student assessment. Fig. 2 depicts the deployment phase of a serious game using our process, once questionnaires are no longer required.

4. Case studies

We summarize two case-studies that we have carried out to test the previous process with two different serious games. For each case-study, we briefly describe the game goal and main mechanics, the interaction data captured (corresponding to the step of the process described in Section 3.1), the analysis of those data traces to derive GLA variables (corresponding to Section 3.2), and the prediction models created to assess players automatically based on their game interactions (Section 3.3) in the consequent deployment phase. Additionally, the case studies have different target variables to test the full process when predicting after-game performance (case study 1) or increase in the measured characteristic (case study 2).

4.1. First Aid Game, a serious game to teach first aid techniques

The First Aid Game is a serious game developed to teach first aid maneuvers to teenagers (Marchiori et al., 2012). The game presents three different levels, each one depicting a different medical emergency. In each level, players can choose among several courses of action (presented as textual or visual options to select from) to assist the in-game character during the emergency. Multiple interactions with the in-game character are available, as are several in-game tools, such as a defibrillator or a smartphone with which (simulated) emergency services can be contacted. After each level is completed, a score for the level is provided as user feedback, based on the errors made and their relevance. Levels can be replayed to improve the score and, consequently, knowledge of its contents.

We collected all relevant interactions in the game using the verbs and activity-types of the xAPI-SG standard: interactions with in-game elements (character and items) were traced with *interacted* traces with the trace object corresponding to the specific element; selections in multiple-choice situations and questions were collected as *selected* traces with the corresponding object and the result indicating whether the response was correct or not; *initialized* and *completed* traces were used to track start and end times of the game and each level, and scores in game levels were tracked in the result extension of the *completed* traces of the corresponding level.

The xAPI-SG traces were then analyzed to derive GLA variables. Some variables were directly obtained following the xAPI-SG based analysis given in Table 1: game completion, the number of interactions with specific in-game elements, and whether specific multiple-choice questions or situations were failed or not. We defined additional variables based on the game designers' decision, as stated in the LAM, that levels could be replayed: first and maximum scores achieved in each game level, and the number of times that each level was repeated.

Next, we used the GLA variables to predict knowledge of first aid techniques (the ones covered in the game) after playing (Alonso-Fernández et al., 2020), both as a binary pass/fail category, and as the exact knowledge score. The First Aid Game had already been validated in a previous experiment, so we knew the game was effective. We tested different prediction models taking as input the GLA data. Highly accurate results were obtained predicting after-game performance as given in the post-questionnaire. The prediction model that provided the best results was a logistic regression (achieving 90% precision, 98% recall, and 10% misclassification rate), which also allowed the results to be interpreted, and provided a measure of the predictive relevance of each variable. The two most relevant variables turned out to be the scores in the first level played, and the total number of interactions with the in-game character. Based on the LAM, the relevance of these two variables seems to be due to the specific strategies followed by players when playing (trial-and-error, exploratory, …), and their overall engagement during the game – while final playthrough scores were surprisingly not that useful when predicting actual learning.

4.2. Conectado, a serious game to raise bullying and cyberbullying awareness

Conectado is a serious game created to raise awareness about bullying and cyberbullying (Calvo-Morata et al., 2020). The game presents a narrative story where players take on the role of a student during the first week at a new high school. Players then experience a bullying and cyberbullying situation in first person during five in-game days. Players can interact with different in-game characters (classmates, teacher, and parents) and in-game objects, including several electronical devices (mobile phone, computer) on which cyberbullying takes place. Some player's choices, related to talking to the parents and/or teachers about the situation, determine which of the three endings is reached.

Again, we captured relevant interactions following the xAPI-SG format, whose verbs and activity-types sufficed to represent them: *interacted* traces were used to collect all relevant interactions with the game characters and items; *selected* traces tracked choices made in decisions and conversations with other characters; *accessed* traces were used to track scene changes; and *initialized* and *completed* traces informed about the starts and ends of the different game days. The ending reached was encoded in the result extension of the *completed* trace for the full game.

From these xAPI-SG traces, we again derived a set of GLA variables, following the xAPI-SG based analysis of Table 1: count of interactions with each character and item, times each of the game areas was accessed, total time spent in each part of the game, and specific decisions in the game, such as conversational choices and those affecting the ending. We also used an additional variable based on the game LAM: the specific ending reached.

The final set of GLA variables was then used as input for models to predict an increase in bullying and cyberbullying awareness due to playing Conectado (Alonso-Fernández et al., 2020), given by the difference in pre-post questionnaires scores. The prediction model that obtained the best results was a Bayesian regression model (achieving 0,54 mean absolute error with 0,053 standard deviation, both normalized to scale [0–10]), which also provided insights into the influence of specific variables on the predictions. In this case, the most significant variables were those representing the specific in-game ending reached, as well as two variables that reported on the time spent in two in-game days. Spending time in one of these days was correlated with greater awareness, while spending additional time in the other exhibited a negative correlation. Again, we examined the design of the game to understand these results. A better ending reached predicted higher awareness increase, indicating that those players who examined a better behavior in the game also learned more. The positive-correlation day was partially spent dealing with an in-game episode of identity theft. At the same time, the negative-correlation one included an interaction where players could attempt to defend themselves against cyberbullying in a simulated in-game social network. However, since reactions to any player comments were scripted, spending time in the in-game social network acted as an indicator of players becoming distracted away from the main plot. These results showed the importance of aligning the game content to the target group age and interests, something that is especially relevant in a game about (misuse of) technology targeted at technology-savvy teenagers.

5. Discussion

The proposed evidence-based process for assessing players using serious games is based on collecting in-game interaction data, as well as evaluation questionnaire results during the game validation to create the prediction models. We have provided a set of guidelines for the critical steps of choosing interaction data to collect and of deriving useful GLA variables from that data. This process is significantly simplified by using a standard format to represent the data (e.g. the xAPI-SG Profile), as a good baseline set of variables with game learning analytics information can be easily derived from interaction traces using this standard (see Table 1). We have additionally provided a set of analyses and visualizations to be applied to any given set of xAPI-SG traces for further insight using our trace monitor, T-Mon, which is freely and openly available at our GitHub repository. T-Mon also allows exploratory analysis using well-known data-science tools, allowing GLA variable candidates to be identified and evaluated. Additional game-specific variables may be more informative but require in-depth knowledge of the game, or even better, access to the game's LAM.

During the game validation process, both pre-post questionnaires and interacted data are collected, so the prediction models can be created and validated. Note that, although in our approach we have focused on the commonly used pre-post questionnaires, other instruments could also serve as long as they are validated. The GLA information obtained could also serve to model players, improve the game's evaluation process, and adapt it to users' characteristics; even while students are still playing, the partial information collected could be used to adapt the game to players' progress and needs. With the information extracted from GLA data, a deeper insight can be gained about the game design (Loh et al., 2015). The analysis of such data could also support the design of more effective learning environments (Liu et al., 2019). Game validations could also be improved as, for instance, the interaction data could discover players not taking the questionnaires seriously and, this way, such questionnaires could and should be discarded from the validation process (Rowley, 2014).

We consider that the integration of automatic non-intrusive player assessment in serious games can greatly simplify and increase their application in education. Game-based assessment provides multiple benefits but still presents several limitations and drawbacks, including the use of immersion-breaking questionnaires to verify whether students are learning or not. Since the step of formally validating a game is essential before deployment to prove that it works as intended, our process is based on re-using data from the validation step to create valid and accurate prediction models for assessment. Once built, those models can be used during the deployment phase for game-based assessment, providing an automatic means of predicting students' knowledge using data science techniques. This way, the actual large-scale deployment of serious games in real settings can be greatly simplified, as game-based assessments can be obtained without the need to conduct pre-post questionnaires. The predicted knowledge can then be used directly for assessment, or as an additional data-point for the teachers or trainers.

5.1. Limitations

Our process has some limitations. A first limitation is that GLA variables created solely from traces represented using the xAPI-SG Profile (see Table 1) may ignore important game-specific details that could lead to more accurate predictions. However, we consider that these variables provide a good baseline on the information that can be extracted from any SG. Additionally, if accuracy of predictions is not deemed to be high enough, re-analysis of the traces to build better GLA variables is certainly possible – and we have developed T-Mon to make such re-analysis much more accessible. Note that it is possible to reuse the non-profile specific aspects of T-

Mon as a starting point to process other statements that are compliant with the xAPI standard.

A second limitation is that prediction models must be created ad-hoc for each serious game based on their formal validation process; and are only valid for players that are sufficiently similar to those it was validated with. Although this means that our process is not generalizable to all kinds of serious games, we consider that it is generalizable to games that share the same genre or mechanics, as similar games can be expected to report similar interaction data, amenable to similar analysis and likely to yield similar results. In a related context, some authors have pointed out that the choice of variables has a greater impact than that of prediction models (Gardner & Brooks, 2018); therefore, the baseline set of variables proposed following the xAPI-SG standard, possibly with additional game-dependent variables, is expected to yield accurate-enough results. Quality of results will be tightly linked to the quality of the games and the selection of variables, which is in turn driven by their Learning Analytics Models, which define how the game and learning designs map to collected interaction data. To ensure the validity of the whole process, it is essential that all games undergo the validation step where prediction models are created, and their validity is tested.

5.2. Conclusions

Serious games have proven to be a useful tool for learning. Availability of a systematic and generalized processes to assess their players would greatly increase serious games' applicability and adoption in real settings, such as schools. It is also time for teachers, educators, and institutions to start trusting these educational tools for assessment purposes. Embedding automatic non-intrusive assessment directly into the game experience provides several advantages, such as reducing both costs and students' stress towards paper-based formal evaluations, while also reducing costs in terms of both time and effort for teachers. Research on automatic assessment using machine learning techniques is been conducted in similar fields, for instance, in language learning through gamified applications (Settles and Laflair, 2020).

Our evidence-based process to assess players using serious games based on tracked in-game interaction data can provide a general standards-based process for other assessment. First, a standard format such as the xAPI-SG Profile can represent the most common interactions present in serious games. Then, their analysis to yield a default set of GLA variables as described on this paper, including their refinement in T-Mon starting from the default analysis and visualizations, provides a baseline of the information that can be extracted from any serious game. With the resulting GLA variables, interpretable prediction models (Carvalho et al., 2019) can be tested during the game validation phase and, in the deployment phase, used to assess players. If possible, we recommend building a Learning Analytics Model early on during the game design, thus ensuring that representative data will be captured, and will therefore be available for later analysis, ready to yield information with educational value (Ke & Shute, 2015). Availability of a LAM can then be used to inform better choices for GLA variables and build better player models or to adapt and personalize games based on the characteristics of their players.

The full lifecycle of serious games is considered in our proposal. Games undergo a first validation phase to prove their efficacy as learning tools, while game learning analytics data is also collected. At the end of this phase, prediction models are created and validated based on in-game interactions and external pre-post questionnaires. When games move to the deployment phase in actual educational settings, the prediction models created can be used to provide a (predicted) score as assessment for each student/player. This way, the cost and time required to deploy the games is greatly reduced, as questionnaires are no longer required to assess students. Use of our process also greatly simplifies large-scale deployment of serious games in real settings and provides a deep insight into the learning experiences of its players.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially funded by Regional Government of Madrid (eMadrid S2018/TCS4307, co-funded by the European Structural Funds FSE and FEDER), by the Ministry of Education (TIN2017-89238-R), by the European Commission (Erasmus+ IMPRESS 2017-1-NL01-KA203-035259), by MIT-La Caixa (MISTI program, LCF/PR/MIT19/5184001) and by Telefonica-Complutense Chair on Digital Education and Serious Games. Also thanks to Julio Santilario Berthilier and Ana Rus Cano for their contribution to Conectado and T-Mon.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2019. Applications of data science to game learning analytics data: A systematic literature review. Computers & Education 141. https://doi.org/10.1016/j.compedu.2019.103612.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2020. Evidence-based evaluation of a serious game to increase bullying awareness. Interactive Learning Environments 1–11. https://doi.org/10.1080/10494820.2020.1799031.

Alonso-Fernández, C., Cano, A.R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2019. Lessons learned applying learning analytics to assess serious games. Comput. Hum. Behav. 99, 301–309. https://doi.org/10.1016/j.chb.2019.05.036.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., Fernández-Manjón, B., 2020. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. Journal of Computer Assisted Learning 36 (3), 350–358. https://doi.org/10.1111/jcal.12405.
Boada, I., Rodriguez-Benitez, A., Garcia-Gonzalez, J.M., Thió-Henestrosa, S., Sbert, M., 2016. 30: 2: a game designed to promote the cardiopulmonary resuscitation protocol. Int. J. Comput. Games Technol. 2016, 1–14. https://doi.org/10.1155/2016/8251461.
Brinton, C.G., Buccapatnam, S., Chiang, M., Poor, H.V., 2016. Mining MOOC clickstreams: video-watching behavior vs. in-video quiz performance. IEEE Trans. Signal Process. https://doi.org/10.1109/TSP.2016.2546228.
Calderón, A., Ruiz, M., 2015. A systematic literature review on serious games evaluation: an application to software project management. Comput. Educ. 87, 396–422. https://doi.org/10.1016/j.compedu.2015.07.011.
Calvo-Morata, A., Rotaru, D.C., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2020. Validation of a Cyberbullying Serious Game Using Game Analytics. IEEE Transactions on Learning Technologies 13 (1), 186–197. https://doi.org/10.1109/TLT.2018.2879354.
Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: a survey on methods and metrics. Electronics 8 (8), 832. https://doi.org/10.3390/electronics8080832.
Charleer, S., Moere, A.V., Klerkx, J., Verbert, K., De Laet, T., 2018. Learning analytics dashboards to support adviser-student dialogue. IEEE Trans. Learning Technol. 389–399. https://doi.org/10.1109/TLT.2017.2720670.
Dicerbo, K.E., 2013. Game-based assessment of persistence. Educ. Technol. Soc. 17 (1), 17–28.
Donker, T., Van Esveld, S., Fischer, N., Van Straten, A., 2018. 0Phobia – towards a virtual cure for acrophobia: study protocol for a randomized controlled trial. Trials. https://doi.org/10.1186/s13063-018-2704-6.
Dörner, R., Göbel, S., Effelsberg, W., & Wiemeyer, J. (Eds.), 2016. Serious Games. Cham: Springer International Publishing. DOI:10.1007/978-3-319-40612-1.
El-Nasr, M., Drachen, A., & Canossa, A., 2013. Game Analytics: Maximizing the Value of Player Data. in: Seif El-Nasr, M., Drachen, A., & Canossa, A., (eds.). London: Springer London. DOI:10.1007/978-1-4471-4769-5.
European Commission, 2018. 2018 reform of EU data protection rules. Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.
Freire, M., Serrano-Laguna, A., Manero-Iglesias, B., Martínez-Ortiz, I., Moreno-Ger, P., Fernánez-Manjón, B., 2016. Game Learning Analytics: Learning Analytics for Serious Games. Learning, Design, and Technology 1–29. https://doi.org/10.1007/978-3-319-17727-4_21-1.
Gardner, J., Brooks, C., 2018. Student success prediction in MOOCs. User Model User-Adap Inter. https://doi.org/10.1007/s11257-018-9203-z.
Gašević, D., Dawson, S., & Siemens, G., 2015. Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64–71. DOI:10.1007/s11528-014-0822-x.
Halverson, R., Owen, V.E., 2014. Game-based assessment: an integrated model for capturing evidence of learning in play. IJLT 9 (2), 111. https://doi.org/10.1504/IJLT.2014.064489.
Homer, B. D., Ober, T. M., & Plass, J. L., 2018. Digital Games as Tools for Embedded Assessment. in: The Cambridge Handbook of Instructional Feedback (pp. 357–375). Cambridge University Press. DOI:10.1017/9781316832134.018.
Ifenthaler, D., Eseryel, D., & Ge, X., 2012. Assessment for Game-Based Learning. In Assessment in Game-Based Learning (pp. 1–8). New York, NY: Springer New York. DOI:10.1007/978-1-4614-3546-4_1.
Jupyter Team, 2020. Jupyter Projects. Retrieved November 1, 2020, from https://jupyter.readthedocs.io/en/latest/projects/content-projects.html.
Kato, P.M., Klerk, S. De., 2017. Serious games for assessment: welcome to the jungle. J. Appl. Test. Technol. 18, 1–6.
Ke, F., & Shute, V. J., 2015. Serious Games Analytics. in: Loh, C. S., Sheng, Y., & Ifenthaler, D., (eds.), Serious Games Analytics. Cham: Springer International Publishing. DOI:10.1007/978-3-319-05834-4.
Kim, Y. J., Ruipérez-Valiente, J. A., Tan, P., Rosenheck, L., & Klopfer, E., 2019. Towards a Process to Integrate Learning Analytics and Evidence-Centered Design for Game-based Assessment. in: Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, (March), 8–10.
Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J., 2017. Learning Analytics as an Assessment Tool in Serious Games: A Review of Literature. in: Serious Games and Edutainment Applications (pp. 537–563). Cham: Springer International Publishing. DOI:10.1007/978-3-319-51645-5_24.
Liu, M., Li, C., Pan, Z., & Pan, X., 2019. Mining big data to help make informed decisions for designing effective digital educational games. Interactive Learning Environments. DOI:10.1080/10494820.2019.1639061.
Loh, C. S., & Sheng, Y., 2015. Measuring Expert Performance for Serious Games Analytics: From Data to Insights. in: Serious Games Analytics (pp. 101–134). Cham: Springer International Publishing. DOI:10.1007/978-3-319-05834-4_5.
Loh, C. S., Sheng, Y., & Ifenthaler, D., 2015. Serious Games Analytics: Theoretical Framework. in: Serious Games Analytics (pp. 3–29). Cham: Springer International Publishing. DOI:10.1007/978-3-319-05834-4_1.
Long, P., Siemens, G., Gráinne, C., & Gašević, D., 2011. LAK '11: proceedings of the 1st International Conference on Learning Analytics and Knowledge, February 27 - March 1, 2011, Banff, Alberta, Canada. in: 1st International Conference on Learning Analytics and Knowledge (p. 195). Retrieved from https://dl.acm.org/citation.cfm?id=2090116.
Michael, D. R., & Chen, S. L., 2005. Serious Games: Games That Educate, Train, and Inform. Education, October 31, 1–95. DOI:10.1145/2465085.2465091.
Project Jupyter, 2020. Jupyter. Retrieved November 1, 2020, from https://jupyter.org/.
Rowley, J., 2014. Designing and using research questionnaires. Management Research Review, 37(3), 308–330. DOI:10.1108/MRR-02-2013-0027.
Ruipérez-Valiente, J. A., Cobos, R., Muñoz-Merino, P. J., Andujar, Á., & Delgado Kloos, C., 2017. Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC. in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 263–272). DOI:10.1007/978-3-319-59044-8_31.
Marchiori, E.J., Ferrer, G., Fernández-Manjón, B., Povar-Marco, J., Suberviola, J.F., Gimenez-Valverde, A., 2012. Video-game instruction in basic life support maneuvers. Emergencias 24 (6), 433–437.
Pérez-Colado, I., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2018. Game learning analytics is not informagic! 2018 IEEE Global Engineering Education Conference (EDUCON) 1729–1737. https://doi.org/10.1109/EDUCON.2018.8363443.
Serrano-Laguna, Á., Manero, B., Freire, M., Fernández-Manjón, B., 2017. A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes. Multimed. Tools Appl. 77 (2), 2849–2871. https://doi.org/10.1007/s11042-017-4467-6.
Settles, B., & Laflair, G. T., 2020. Machine Learning Driven Language Assessment, 8, 247–263.
Shoukry, L., Göbel, S., & Steinmetz, R., 2014. Learning Analytics and Serious Games: Trends and Considerations. in: Proceedings of the 2014 ACM International Workshop on Serious Games. DOI:10.1145/2656719.2656729.
Shute, V., Ke, F., & Wang, L., 2017. Assessment and Adaptation in Games. in: Instructional Techniques to Facilitate Learning and Motivation of Serious Games (pp. 59–78). Cham: Springer International Publishing. DOI:10.1007/978-3-319-39298-1_4.
Serrano-Laguna, A., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., Fernández-Manjón, B., 2017. Applying standards to systematize learning analytics in serious games. Computer Standards & Interfaces 50, 116–123. https://doi.org/10.1016/j.csi.2016.09.014.
Shute, V.J., Rahimi, S., 2017. Review of computer-based assessment for learning in elementary and secondary education: computer-based assessment for learning. J. Comput. Assisted Learning 33 (1), 1–19. https://doi.org/10.1111/jcal.12172.
Shute, Valerie J., Ventura, Matthew, Kim, Yoon Jeon, 2013. Assessment and learning of qualitative physics in newton's playground. J. Educ. Res. 106 (6), 423–430. https://doi.org/10.1080/00220671.2013.832970.
Sliney, A., & Murphy, D., 2011. Using Serious Games for Assessment. in: Serious Games and Edutainment Applications (pp. 225–243). London: Springer London. DOI: 10.1007/978-1-4471-2161-9_12.
Devin, Soni, 2007. Supervised vs. Unsupervised Learning •. Retrieved from https://www.kdnuggets.com/2018/04/supervised-vs-unsupervised-learning.html.
Wallner, G., & Kriglstein, S., 2015. Comparative Visualization of Player Behavior for Serious Game Analytics. in: Serious Games Analytics (pp. 159–179). Cham: Springer International Publishing. DOI:10.1007/978-3-319-05834-4_7.

6.1.5. Lessons learned applying learning analytics to assess serious games

Full citation

Cristina Alonso-Fernández, Ana Rus Cano, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2019): **Lessons learned applying learning analytics to assess serious games.** Computers in Human Behavior, Volume 99, October 2019, Pages 301-309. DOI: 10.1016/j.chb.2019.05.036.

Impact metrics: JCR 2019 Impact Factor: 5.003, Q1 in Psychology, Experimental.

Abstract

Serious Games have already proved their advantages in different educational environments. Combining them with Game Learning Analytics can further improve the life-cycle of serious games, by informing decisions that shorten development time and reduce development iterations while improving their impact, therefore fostering their adoption. Game Learning Analytics is an evidence-based methodology based on in-game user interaction data, and can provide insight about the game-based educational experience promoting aspects such as a better assessment of the learning process. In this article, we review our experiences and results applying Game Learning Analytics for serious games in three different scenarios: (1) validating and deploying a game to raise awareness about cyberbullying, (2) validating the design of a game to improve independent living of users with intellectual disabilities and (3) improving the evaluation of a game on first aid techniques. These experiences show different uses of game learning analytics in the context of serious games to improve their design, evaluation and deployment processes. Building up from these experiences, we discuss the results obtained and provide lessons learnt from these different applications, to provide an approach that can be generalized to improve the design and application of a wide range of serious games in different educational settings.

Full publication

Computers in Human Behavior 99 (2019) 301–309

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Lessons learned applying learning analytics to assess serious games

Cristina Alonso-Fernández*, Ana R. Cano, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón

Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, C/ Professor José García Santesmases, 9. 28040, Madrid, Spain

ARTICLE INFO

Keywords:
Learning analytics
Game analytics
Serious games
Game-based learning
Evidence-based learning*Subject classification codes:*
Analysis and evaluation methods
Case studies
Design experiments
Surveys and questionnaires
Learning technologies and tools

ABSTRACT

Serious Games have already proved their advantages in different educational environments. Combining them with Game Learning Analytics can further improve the life-cycle of serious games, by informing decisions that shorten development time and reduce development iterations while improving their impact, therefore fostering their adoption. Game Learning Analytics is an evidence-based methodology based on in-game user interaction data, and can provide insight about the game-based educational experience promoting aspects such as a better assessment of the learning process. In this article, we review our experiences and results applying Game Learning Analytics for serious games in three different scenarios: (1) validating and deploying a game to raise awareness about cyberbullying, (2) validating the design of a game to improve independent living of users with intellectual disabilities and (3) improving the evaluation of a game on first aid techniques. These experiences show different uses of game learning analytics in the context of serious games to improve their design, evaluation and deployment processes. Building up from these experiences, we discuss the results obtained and provide lessons learnt from these different applications, to provide an approach that can be generalized to improve the design and application of a wide range of serious games in different educational settings.

1. Introduction

The gaming industry has experienced a vast growth worldwide in recent years (Entertainment Software Association, 2017). The application of games with a non-entertainment primary purpose, so-called Serious Games (SGs) (Abt, 1970), can provide multiple benefits (Boyle et al., 2016) in environments where games were not traditionally present. Such is the case of education, where non-interactive contents still constitute the majority of learning materials, and there is little consensus about how to best include technology in the classrooms (Adams Becker et al., 2017). However, this slow adoption of learning games in a broad sense, contrasts with the use of games in specific educational domains (e.g. business, military (Kato & Klerk, 2017)) and with the major presence of games in the spare time of students (Pew Research Center, 2018).

New techniques such as Learning Analytics (LA), are trying to provide insight about the educational processes and improve the common educational scenarios benefiting from data-driven approaches. LA, as defined in (Long, Siemens, Gráinne, & Gašević, 2011), aims to measure, collect, analyze and report data from learning tools, such as LMSs (Learning Management Systems) or MOOCs (Massive Open Online Courses), to extract useful information about how students learn with the purpose of understanding and optimizing their learning processes and contexts (Sclater, 2017).

LA techniques can clearly be applied to game environments, where their interactive nature is adequate to the data-capturing process. Data from serious games can therefore be collected while students are playing – both providing information about the impact the game is making (e.g. in their learning, as will be the goal of LA) but also providing information about the appropriateness of the game design and its mechanics (aligning with the Game Analytics field for entertainment games (Seif El-Nasr, Drachen, & Canossa, 2013)). This combination of LA and GA techniques results in Game Learning Analytics (GLA) (Freire et al., 2016) for serious games. Consequently, in-game interactions can be analyzed with many different purposes, in particular: 1) to better understand how students learn using games and 2) to validate the actual educational and game designs. Several studies have been carried out for these two purposes for instance: for purpose 1), authors have been able to assess students based on in-game data (Kiili, Moeller, & Ninaus, 2018) or to predict learning results based on students' interactivity (Hernández-Lara, Perera-Lluna, & Serradell-López, 2019); while for purpose 2), learning analytics data have been used to validate

* Corresponding author.
E-mail addresses: calonsofernandez@ucm.es, crisal03@ucm.es (C. Alonso-Fernández), anarcano@ucm.es (A.R. Cano), acmorata@ucm.es (A. Calvo-Morata), manuel.freire@fdi.ucm.es (M. Freire), imartinez@fdi.ucm.es (I. Martínez-Ortiz), balta@fdi.ucm.es (B. Fernández-Manjón).

https://doi.org/10.1016/j.chb.2019.05.036
Received 1 August 2018; Received in revised form 3 May 2019; Accepted 29 May 2019
Available online 01 June 2019
0747-5632/

serious games (Tlili, Essalmi, Jemni, & Kinshuk, 2016) or to find possible improvements in game design (Hicks et al., 2016).

The application of analytics to serious games is not new, however, few studies have reported empirical evidence to inform about the learning process adequately (Chaudy, Connolly, & Hainey, 2014). Despite the increased interest in the application of LA techniques for assessment in serious games, authors have identified the need for more data-based research regarding this topic (Liu, Kang, Liu, Zou, & Hodson, 2017).

This papers' contribution aims to fill the gap found in the literature available by providing data-based evidence of the possible applications of GLA data for serious games, giving an example of three specific experiences. Each of these three application scenarios were performed with a different SG, was focused on a different domain and was used in real-world educational scenarios, instead of only relying on experiments in controlled environments. The SG *Conectado* is a tool to be used by the teachers in the classroom to address bullying and cyberbullying (Calvo-Morata, Rotaru, Alonso-Fernandez, Freire, Martinez-Ortiz, & Fernandez-Manjon, 2018). GLA was used in *Conectado* to improve validation and deployment in schools. The SG *DownTown* is designed for promoting independence in users with intellectual disabilities (Cano, Fernández-Manjón, & García-Tejedor, 2018); and its use of GLA illustrates how to validate a game design in situations where information cannot be directly gathered from the users. Finally, the *First Aid Game* is designed to teach first-aid maneuvers to teenagers, and had already been formally validated (Marchiori et al., 2012). The use of GLA for the *First Aid Game* focuses on improving the evaluation and deployment of games by applying data mining models to predict students' knowledge after playing based on interaction data (Alonso-Fernández, Caballero Roldán, Freire, Martinez-ortiz, & Fernández-Manjón, 2019), proving that games can help to accurately assess students' knowledge. This can greatly contribute to SG generalization and adoption in schools.

These three games have been chosen to showcase the usefulness of GLA, as they are representatives of different goals: *Conectado* aims to raise awareness; *First Aid Game* aims to improve students' knowledge; and *DownTown* aims to train skills. These case studies also have different target users: *DownTown* focuses on adults with intellectual disabilities, like Down Syndrome or Autistic Spectrum Disorders; while the other two studies focus on young students, although while *First Aid Game* is designed focusing only on the players, *Conectado* is designed as a tool to be used in class supervised by teachers. Due to the different goals, target users and educational scenarios, GLA played a different role during lifecycle of the serious game development: from game validation (*DownTown*), to overcome deployment issues in actual scenarios (*Conectado*) and students' assessment (*First Aid Game*).

This paper summarizes our experiences using GLA to provide teachers and researchers with reliable, evidence-based insights on the accuracy of the game design to the expected learning outcomes, and to facilitate their deployment in actual real educational scenarios. Therefore, we consider that this GLA approach can be generalized to improve the design and application of a wide range of SGs in different educational settings. The paper is structured as follows: Section 2 describes the methodology followed while applying GLA techniques; the three case studies collected are explained in detail in Sections 3, 4 and 5; finally, Section 6 discusses the results and Section 7 summarizes the main conclusions of the different applications of GLA data for Serious Games.

2. Methodology

In this paper, we review three different case studies applying game learning analytics data with serious games. Each of the three games were developed based on specific educational designs, which established the different goals to be achieved with each game. To connect these goals to the actual Learning Analytics data to be analyzed, we followed the Learning Analytics Model (LAM) (Perez-Colado, Alonso-Fernández, Freire-Moran, Martinez-Ortiz, & Fernández-Manjón, 2018). LAMs describe: 1) how the educational design and learning goals are linked with specific game goals and game mechanics and 2) which interaction data should be gathered, how it will be collected, and how to analyze that data to be meaningfully presented to the different stakeholders. Traditionally, collected data is analyzed after the gameplay session is over as a report for the current gameplay session, to generate aggregated reports from several game plays or to extract complex metrics and relationships (high computational cost). However, we consider SGs as an educational tool that can be used also during the class, so it is required to provide some feedback/insights while games are running using near real-time analysis. In both scenarios, the results can be used to fill dashboards that provide visual feedback (at near real-time or for later analysis) about the performance of the students for the involved stakeholders, such as teachers, students, or managers of educational institutions.

The three games reviewed in this paper had different learning goals, and their development was driven by their LAMs in order to extract the data of interest considering the specific goals and design characteristics of each game. Once GLA data is collected, it can be analyzed to validate or refute the appropriateness of the educational design of each game. For instance, checking that the game mechanics are properly designed for its target users in terms of complexity, duration, number of tasks assigned, etc. Results of the analysis can also be used to validate additional hypothesis established by educators or researchers regarding the expected learning outcomes and abilities of the players while interacting with the game.

The GLA data collected in all three scenarios, specified in their LAMs, followed the xAPI-SG Profile (Serrano-Laguna et al., 2017), a standard collection model for tracking interaction data from Serious Games. This interaction model is implemented in Experience API (xAPI) (ADL, 2012), a data format to track learning activities whose traces have three main fields: the "actor" who makes the action, a "verb" which is the action itself and an "object" which is the target of the action. This tracking model is similar to the model used to track user activity in social networks defining a common set of verbs, activity types and extensions.

All xAPI-SG traces collected (see examples depicted in the figures in Subsections 3.2., 4.2. and 5.2.) use a unique pseudo-anonymous token given to players as name; describe the interactions using a combination of in-game actions and targets, possibly including additional data in the "results" field, and contain a timestamp identifying the time in which the interactions occurred.

Of course, enjoying the advantages of collecting GLA data comes at a price, and in this case a GLA infrastructure is required to be able to collect and analyze the data. In our case, we have reused both the xAPI tracker and the GLA open code infrastructure developed in the H2020 projects RAGE and BEACONING that is available online[1].

Each of the following three sections begins by providing an overview of the goals, target users and design of the SGs used in the respective case study, including their experimental designs, the evaluation methods, the GLA data collected, and finally, the results obtained via analysis.

3. Case study: *Conectado*

Conectado is a graphic adventure SG to increase bullying and cyberbullying awareness for students between 12 and 17 years old, which is considered a serious universal problem (Kowalski, Giumetti, Schroeder, & Lattanner, 2014). The game places the player in the role of a student suffering cyberbullying by schoolmates upon arriving in a new school. The aim of the game is to promote empathy with the victims making players experience the feelings that aggression victims

[1] https://github.com/e-ucm/rage-analytics.

Fig. 1. Screenshots of *Conectado* depicting the classroom scenario and the home use of the in-game mobile classmates social chat.

```
{
  "actor" : {
     "name" : "XXXX"
  },
  "verb" : {
     "id" : "http://adlnet.gov/expapi/verbs/interacted"
  },
  "object" : {
     "id" : "http://a2:3000/api/proxy/gleaner/games/<game-id>/<version-id>/Computer",
     "definition" : {
        "type" : "https://w3id.org/xapi/seriousgames/activity-types/game-object",
     },
  "result" : {
      "extensions" : {
        "GameDay" : 1.0,
        "GameHour" : "21:30",
        "MobileMessages" : "True"
    }
  },
  "timestamp" : "2018-05-17T12:04:56.835Z"
}
```

Fig. 2. Example collected xAPI-SG trace, describing a player's interaction with the in-game computer in *Conectado*.

usually experience. The game is designed to be played at schools with the supervision of teachers; the game is therefore intended to be completed in 30–40 min, leaving time for a 15-min discussion with the teacher afterward, which would fit into a standard 55-min high school lecture in Spain.

The plot of the story occurs during 5 days in which the players move from home to school and back home, in the usual places where cyberbullying occurs (Fig. 1). Empathy and other emotions are brought to players through mini-games presented as nightmares at the end of each day. Dialogues with other in-game characters also raise players' emotions, as the schoolmates slowly turn against the main character.

The player can choose among several options in different in-game situations. Choices taken alter the story, e.g. the ending is determined by the protagonist's relationship with classmates and parents, and by whether the character has asked the teacher for help or not. No matter the choices taken, a satisfactory ending cannot be reached until the end of the fifth day, so all players go through the complete game experience. Also, physically aggressive answers are excluded from dialogues.

3.1. Experimental design

Conectado provides a linear flow with some available choices to take in-game that arrive in one of the three possible endings. The game comprises the most common situations, scenarios and roles involved in bullying and cyberbullying, as identified in literature (El Asam & Samara, 2016; Larrañaga, Yubero, Ovejero, & Navarro, 2016; Patchin & Hinduja, 2006). Evaluation of the game is done with pre- and post-tests which assess the level of cyberbullying awareness. The questionnaire used in the pre- and post-tests was adapted based on previous (cyber) bullying measurement questionnaires (Álvarez-García, Núñez Pérez, & Dobarro González, 2013; Ortega-Ruiz, Del Rey, & Casas, 2016). Reliability of questionnaire was verified statistically (Cronbach's Alpha = 0.95). Full questionnaire can be found (in Spanish) as an Appendix in (Calvo Morata, 2017).

The game was validated in a single group pre- and post-test experiment (reasons for this choosing are fully explained in the experiment publication) (Calvo Morata, 2017) with N = 257 high-school students aged between 12 and 17 years old from three schools in Spain, in June 2017. The pre- and post-test shared 18 questions to measure bullying and cyberbullying awareness.

3.2. GLA data

GLA data collected included: options taken in some relevant dialogues (e.g. if players have shared their personal password or not, if they have decided to tell parents or teachers about the situation, if they have chosen the most confronting option when meeting the aggressor), changes in the patterns of friendship with the classmates and parents and the general risk value, the specific ending reached (out of the 3 possible), interactions with other classmates and parents, interactions with other game elements (e.g. mobile phone, computer, school bathroom), times in completing each game day and the full game, and whether players have completed the game or not.

Fig. 2 depicts an example xAPI-SG trace collected for *Conectado* showing that the player with name "XXXX" has interacted (xAPI-SG verb, depicted in red) with the game object (xAPI-SG activity type, in blue) with identifier "Computer" (orange); extensions (green) represent the day and hour in-game and that the protagonist has mobile

messages.

3.3. Results

The increase in cyberbullying awareness was measured with the pre-test and post-test, each containing eighteen 7-point Likert items to validate the game. The average score in the pre-test was 5.72 (SD = 1.26), compared to 6.38 (SD = 1.11) in the post-test, a statistically-significant effect (paired Wilcoxon test yields p-value < 0.001).

In the optional questions of the post-test, out of the players who answered (73.3%), most of them (85.9%) considered that they had learned something new about bullying or cyberbullying while playing, and 80% of the opinions given about the game were positive. Most students (88%) did not feel represented by any in-game character but some admitted feeling identified with the classmates (9%) or the victim (2%).

The analysis of GLA data showed that the younger the players, the longer the time required to complete the game (for 16 years old, mean completion time was 28 min, which increased up to 38 min for 12 years old). Also, women took longer to finish the game than men (36 min on average for female players, compared to 31 min for males). Regarding the three possible endings, data shows that most players who finished arrived at the best ending (74.4%).

3.4. GLA application: lessons learned

GLA data was used at near real-time to allow teachers to monitor what students were doing while they were playing the game (Calvo-Morata, Alonso-Fernández, Freire, Martínez-Ortiz, & Fernández-Manjón, 2018). This information was provided using a dashboard that comprised several visualizations including the ones depicted in Fig. 3: gauge chart showing the average friendship level with in-game characters (a); bar chart showing number of players that were in each day in-game (b); and pie chart showing the players in each possible ending (c).

Results of GLA helped to identify some design problems, for instance, the first version of the game (tested in an initial formal evaluation with N = 64 students) took too long to complete and left no additional time for discussion with the teacher; this was amended in the second version of the game. All the results reported correspond to this second version of the game. Also, as data was captured at near real-time, some available visualizations allowed to have feedback to control the intervention (e.g. check if data was being received correctly, at which stage of the game players were at every moment). The game was validated as data proved that all target users (despite their different age, gender, school, previous intervention conditions) indeed increased their cyberbullying awareness with the intervention and most of them found the game enjoyable.

This experiment has so far been deployed in 8 schools where some of the interventions where performed by instructions with little support by the authors. Hence, this experiment leverages the advantages of GLA to facilitate the deployment of SGs in classroom settings in a systematic way.

4. Case study: *DownTown*

DownTown, a Subway Adventure is an espionage-themed game for players from 18 to 45 years old with Intellectual Disabilities (ID) such as Down Syndrome, mild cognitive disability or certain types of Autism Spectrum Disorder (ASD). The game aims to train them in using the public subway transportation system of Madrid (Spain), following the positive results of other game-learning experiences with ID users (Kwon & Lee, 2016).

The game was developed in a 3D realistic perspective, so players can identify the in-game scenarios with reality when they travel alone, as depicted in Fig. 4. Users can navigate in the game as they do it in real life, so they are trained in choosing the right route travelling from one station to another. Routes are randomly assigned by the game based on the selected difficulty and also can be manually configured. Although specific cognitive skills may vary among individuals, the game design considers the most common cognitive features and barriers of the users. The game includes different levels, so players can progress from easier tasks to more complex ones. Default routes are available, but they can also be manually configured, so players can play the routes that they travel the most in their daily routine.

Downtown also includes puzzles and quests designed to train basic daily skills (e.g. independence, long- and short-term memory, spatial vision) and social aspects that are complex for the target users (e.g. interacting with subway operators) to promote their independent life. Further details about the game design can be found in (Cano, Fernández-Manjón, & García-Tejedor, 2016).

4.1. Experimental design

DownTown provides different missions in four available difficulty levels. The design considers the cognitive needs of the target users which were especially relevant when it comes to develop the game mechanics and procedures. As feedback could not be directly gathered from target users (e.g. questionnaires), the evaluation was fully based on GLA data, meaning that traced interactions provided all the data used to validate the game design.

The game was tested with N = 51 adults with intellectual disabilities (Down Syndrome, mild cognitive Disability or certain types of Autism Spectrum Disorder), aged between 19 and 41 years old, in the Fundación Síndrome de Down in Madrid (Spain), in May–June 2017. Students played a total of 3 h spread over 3 sessions.

4.2. GLA data

GLA data collected included: avatar configuration and accessibility preferences, attempts to complete each minigame, number of correct and incorrect stations in each route, number of clicks in some interface

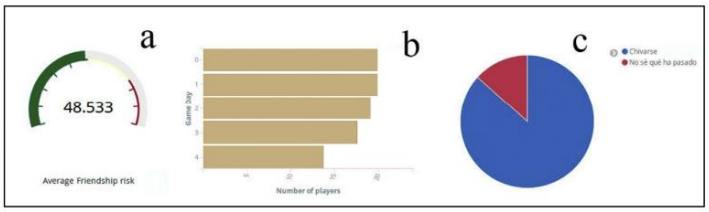

Fig. 3. Some of the visualizations shown in the dashboard for teachers while students were playing *Conectado*: average friendship level of players with in-game characters; day in-game players are at; and ending reached by players.

Fig. 4. Screenshots of *DownTown* depicting the user avatar in a Metro wagon and in a hall before using the ticket.

```
{
  "actor" : {
     "name" : "XXXX"
  },
  "verb" : {
     "id" : "http://adlnet.gov/expapi/verbs/progressed"
  },
  "object" : {
     "id" : "http://a2:3000/api/proxy/gleaner/games/<game-id>/<version-id>/Mission_1_Ex_ElMaletinMarron",
     "definition" : {
        "type" : "https://w3id.org/xapi/seriousgames/activity-types/quest",
     }
  },
  "result" : {
     "extensions" : {
        "https://w3id.org/xapi/seriousgames/extensions/progress" : 0.3333333
     }
  },
  "timestamp" : "2018-01-08T18:28:36.211Z"
}
```

Fig. 5. Example collected xAPI-SG trace, describing a player's progress in a quest in *DownTown*.

elements (e.g. accessibility menu, "help" items), progress per timestamp, time spent in each route, time in completing each session or minigame, total game time, inactivity times, and time and number of attempts completing each task after asking for help.

Fig. 5 depicts an xAPI-SG trace collected showing that the player with name "XXXX" has progressed (xAPI-SG verb, red) 0.33 (extension, green) in the quest (xAPI-SG activity type, blue) with identifier "Mission_1_Ex_ElMaletinMarron" (orange).

4.3. Results

GLA data showed that most students (85.8%) were able to reach their destination following the right path. Half of the mistakes (50.8%) occurred during the first 30 min of playing (once students completed a few routes to understand the mechanics). Beyond making less errors, users also improved their performance in the videogame as they played more sessions reducing their inactivity time (by an average of 58%).

Additional hypotheses were contrasted with data: no significant differences were found between players who customized their avatars (71%) and players who did not (29%) (as suggested by majority of literature (Griebel, 2006; Klimmt, Hefner, Vorderer, Roth, & Blake, 2010; Newman, 2002, pp. 405–422), although it may be significative that the majority of the users that changed the avatar were diagnosed with Down Syndrome); players with previous transportation training completed routes faster (taking 6% less time); and regular videogame players completed the tasks quicker (taking 12% less time), showing less inactivity time and making less mistakes that non-players.

4.4. GLA application: lessons learned

The analysis of the GLA data also helped to validate the game design. From the results, it was drawn that the complexity of the tasks in the different levels was adequately adjusted to the users' intellectual abilities, as there were no peaks in resolution times; that the number of tasks in each level was balanced with its difficulty, as each game level included one more task on average than the level with difficulty immediately below; and that there was a failure in the game design as there was not a correspondence between the number of stations that users navigate in each level and its difficulty (users transited more stations –eleven– on average in the "medium" level than in the "hard" level –ten–, which was not intended in the game design).

Therefore, results of the analysis helped to validate the game design decisions regarding the cognitive skills of the users as well as to identify some problems in the development (that were not found in the beta-testing of the game). Hypotheses that educators had regarding the abilities of the users based on their different IDs and previous experience were contrasted with the data and it has been proven that, after some initial time to become familiar with the game environment, the majority of the users were able to achieve the learning goals of the game (i.e. successfully reach a destination). Next phase of the research will be to compare the behavior of the users that played *DownTown* in the subway versus the ones that were not previously trained with the videogame.

5. Case study: *First Aid Game*

First Aid Game aims to instruct in cardiopulmonary resuscitation (CPR) maneuvers for students between 12 and 17 years old, following the guidelines defined by the European Resuscitation Council (ERC) (European Resuscitation Council Guidelines Writing Group, 2015). The game-like simulation presents three different scenarios with an in-game character, as depicted in Fig. 6, suffering from different emergencies (chest pain, choking and/or unconsciousness). The player then faces a linear situation showing the adequate procedure. The game is designed to have a maximum gameplay duration of 30 min and players can repeat each level as many times as they want to.

The specific knowledge to be learned in each of the three situations is assessed with different types of multiple-choice questions that players can retry when choosing a wrong option (Fig. 6). The more mistakes the user makes, the less score is given for that situation. After choosing the

Fig. 6. Screenshots of *First Aid Game* depicting a conversation with the user and an in-game selection using pictures.

right option, additional in-game videos show the complete procedures. The player can also interact with some game assets: the main character suffering from the specific situation, a mobile phone to call the emergency services or a semi-automatic external defibrillator.

The game was validated in a usual pre-post experiment with a control group in 2011 with more than 300 students in four secondary schools of Aragon (Spain), as described in (Marchiori et al., 2012). Results showed that, although slightly lower than in the control group (that took part in a theoretical and practical demonstration of the maneuvers by an accredited instructor), the increase in the results from the learning experience was significant in the experimental group which played the game.

5.1. Experimental design

First Aid Game had an evaluation model based on scores given for each of the three in-game situations. The scores decrease based on the errors made in each situation and their relevance. As the game was already validated in a pre-post experiment with a control group, now it is possible to construct prediction models for the post-test results based on GLA interaction data.

The game was tested with N = 227 high-school students aged between 12 and 17 years old from one school in Madrid (Spain) in January–February 2017. The pre- and post-test shared a common part with 15 questions to assess students' knowledge on the basic life-support maneuvers covered in the game.

5.2. GLA data

GLA data collected included: whether players have completed the game or not, total score obtained, first and maximum scores obtained in each of the three levels, interactions with game elements, correct and incorrect answers in questions and how many times each level was repeated.

Fig. 7 depicts an example xAPI-SG trace collected for *First Aid Game* showing that the player with name "XXXX" has selected (verb, in red) the response "112" (result, green) in question (activity-type, blue) with identifier "NumeroEmergencias" (orange) and, as the result "success" is set to "true" (green), the option selected is the correct one.

5.3. Results

We first checked that learning was still significant in this experiment (i.e. reproducibility of previous results): from pre-test, with mean of 8.06 (SD = 2.05), to post-test, with mean of 9.83 (SD = 2.38), the paired sample Wilcoxon Signed-Rank test showed a statistically significant increase ($p < 0.05$). Then, prediction models were carried out to predict post-test scores with two different targets: exact score in scale [0–15] and pass/fail classification (stablishing "pass" in 8 correct answers out of the 15 questions in the post-test). For both targets, some models were developed taking the pre-test and the GLA data as inputs, while other models took only the GLA data, to further avoid the pre-test. Models were compared using 10-fold cross validation.

5.4. GLA application: lessons learned

For pass/fail predictions, decision trees, logistic regression and Naïve Bayes Classifier were tested. The best model with all previous information was a logistic regression which obtained 89% precision, 98% recall and 10% misclassification rate. Without pre-test information, the best model was another logistic regression with slightly worse results (87% precision, 98% recall and 13% misclassification rate).

For score prediction in scale [0–15], regression trees, linear regression and Support Vector Regression (SVR) with non-linear kernels (polynomial, radial basis and sigmoid) were tested. The best prediction model of scores in range [0–15] taking as input the pre-test and the


```
{
  "actor" : {
     "name" : "XXXX"
  },
  "verb" : {
     "id" : "https://w3id.org/xapi/adb/verbs/selected"
  },
  "object" : {
     "id" : "http://a2:3000/api/proxy/gleaner/games/<game-id>/<version-id>/NumeroEmergencias",
     "definition" : {
        "type" : "http://adlnet.gov/expapi/activities/question",
     }
  },
  "result" : {
      "success" : true,
      "response" : "112"
  },
  "timestamp" : "2017-01-27T03:20:25.571Z"
}
```


Fig. 7. Example collected xAPI-SG trace, describing a player's selection in a question in *First Aid Game*.

interaction data was a SVR whose mean error was 1.5 (SD = 1.3). Without pre-test information, the best model was again a SVR whose mean error increased to 1.6 (SD = 1.4).

Additionally, we have determined the features derived from GLA data that are most relevant in the predictions; these include the total number of interactions with the game character and the score in one game level, providing a baseline of in-game actions that can be traced from any SG for assessment purposes. These specific GLA data were related both with the game mechanics and the educational game design of the *First Aid Game*, highlighting that both must be considered to determine which data should be captured from serious games for assessment.

With these preliminary results, we could predict with high accuracy the students' results in the post-test, therefore avoiding the post-test itself. As expected, results are better when simply predicting pass/fail categories than when predicting the exact score, but still good results are obtained for score predictions. Moreover, in some scenarios, it may suffice for teachers to know whether students have acquired enough knowledge to pass or fail the topic. As expected, models taking as inputs both the pre-test and the xAPI-SG GLA data provide better results than models without pre-test information; however, results in models without pre-test are only slightly worse. Therefore, we could also avoid conducting the pre-test and let students simply play the game and, from their interactions, predict their knowledge after playing. Therefore, game deployment in schools is greatly simplified as there is no need of explicit questionnaires to know if students have learned from the game.

6. Discussion

We have summarized three experiences, including their design, analysis and results, with each focusing on a different application of Game Learning Analytics to the corresponding serious game:

Simplify the validation and deployment in schools of a SG that increases bullying and cyberbullying awareness for students (*Conectado*).

Validate the design of a SG that trains students with intellectual disabilities in using the subway without needing explicit feedback (*DownTown*).

Improve the evaluation and deployment of a SG to teach first aid techniques by predicting knowledge after playing to avoid carrying out the post-test (*First Aid Game*).

From the first experience, with *Conectado,* we have been able to validate the game proving that it indeed increases cyberbullying awareness and that it is a helpful tool for students. Data also showed multiple insights into how users played the game, including how they interacted with other characters and the ending reached. Once the game has been validated for its target users using questionnaires and GLA data, its consequent deployment is greatly simplified as it can be used in a larger number of scenarios without the presence of a researcher, without having to further conduct the pre- and post-tests, or even outside of a classroom setting.

From the second experience, with *DownTown,* we have validated the game design in a scenario where obtaining explicit feedback from the users is not always accurate. After users understand how the game works, results show that they effectively improve in completing routes in the game. The game also has a positive effect on players, increasing their motivation and engagement in the learning process. Data also helped to validate the game design and its mechanics and tested whether the specific goals and mechanics are adequately adapted to the users' intellectual characteristics.

From the third experience, with the *First Aid Game*, we have introduced data mining techniques looking for an improvement in the evaluation of players of SGs using the potential of the GLA data collected from in-game interactions. The highly-accurate results obtained suggest that this approach could indeed be applied to avoid the costly pre-post-tests experiments. This may simplify the deployment of SGs and the evaluation of students, as games could simply be played without the additional questionnaires, allowing longer gameplay times, and without requiring the presence of researchers to conduct the tests, as interaction data could be remotely tracked.

In this work, we have revised three experiences we have carried out applying serious games in real settings and applying game learning analytics techniques with different purposes. We consider that our experiences can provide some guidelines for future research on this topic and authors could benefit from some of the lessons learnt:

Standardize GLA data collection: in our three experiences we have used the standard format to collect data from serious games, the xAPI-SG Profile. The use of this standard has simplified collection as we could easily define and match the interactions to be captured for each game with the specific verbs and activity types of the xAPI-SG Profile. The use of a standard has also simplified integration with larger systems - in our case, we have been able to provide real-time information by connecting the data collected from games with the Analytics System. In addition, this standardization would allow real-time analysis easily comparing the interactions of different games (e.g. times of use, learning, completion). Authors had previously identified the need of data collection standards, to compare studies' outcomes and reuse in-game data (Liu et al., 2017; Smith, Blackmore, & Nesbitt, 2015). We highly encourage researchers to use some standard when collecting game learning analytics data from serious games as it can simplify integration with larger systems and even reusability of data when openly shared for research purposes.

Purposes of GLA data: in our experiences, we have showcased how game learning analytics data can be effectively used for different purposes at different stages of the serious games' lifecycle, and specifically: to validate the game design (*DownTown* case study), to validate and simplify deployment of a game (*Conectado* case study), and to simplify assessment of learners with games (*First Aid Game* case study). Additionally, we have shown examples on how game learning analytics data can provide further information about how students played games or how games promote motivation and engagement. Previous research had been carried out for purposes related to these, for instance: one of the focuses of serious games analytics is to improve game design (Loh, Sheng, & Ifenthaler, 2015); while other works focus on *stealth assessment* (Shute, Ke, & Wang, 2017). The collection of purposes described and exemplified on this work can be used as a baseline for further research.

Stakeholders to benefit from GLA use: It is also important to notice that these purposes cover the interest of different stakeholders: for game designers and developers, to simplify validation of their designs; for teachers and educators, to simplify the application of games in their classes, to obtain real-time information while games are in play, and even to assess their students; for students/learners, to be more effectively assessed based on their in-game interactions and to know their progress and statistics themselves. Previous studies had identified that stakeholders, beyond students and teachers, should be considered for issues related with serious games application for education, as each stakeholder will have their interests and requirements (Jaccard, Hulaas, & Dumont, 2017).

7. Conclusions

On each of the three experiences described on this paper, we have gathered different uses of GLA data for Serious Games: to validate the game design, verifying that design choices were adequate for its goals (e.g. gameplay time, difficulty of levels); to prove that all the game target users can reach the expected outcomes; to test additional hypotheses expected by educators or researchers; to provide visual feedback while games are in play to follow the intervention; or to predict learning results, traditionally measured with questionnaires, simplifying deployment and assessment.

However, these are not the only possible uses of GLA data for Serious Games. At early stages of the design process, interaction data

collected with some target users could help to quickly iterate and find problems in early versions of the game; feedback from users at any stage could be collected remotely simplifying early testing or large deployment; and improvements for subsequent versions of an already deployed game could be extracted from players' interaction data.

From our experience, we have also identified that it is essential that games have an underlying learning design that allows for evidence-based assessment. For this purpose, it is convenient that the design of games follows a LAM to clearly establish its goals, and how they are to be measured with interaction data adequately collected, but this is not enough; from the very beginning, games need to be designed so that data can be extracted from them and provide the information required to validate the games and assess students using them. In that sense, games need to be designed bearing in mind the key data that is to be collected so the desired evaluation has the required input data to be adequately performed. Additionally, it is highly recommended that the collected GLA data follows some standard format.

The positive results obtained in the three experiences reviewed showcase the importance of game learning analytics in different contexts, simplifying serious games' validation and deployment, as well as players' assessment, and even being used as the sole means to obtain players feedback. All these applications intend to promote the use of serious games with different goals in real contexts. We consider that the use of game learning analytics techniques can and should be generalized to improve the design and application of serious games in different educational settings.

Acknowledgments

We would like to thank the anonymous reviewers for their detailed comments and recommendations that have greatly helped us to improve this paper.This work has been partially funded by Regional Government of Madrid (eMadrid P2018/TCS4307), by the Ministry of Education (TIN2017-89238-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676, Erasmus+ IMPRESS 2017-1-NL01-KA203-035259) and by the Telefónica-Complutense Chair on Digital Education and Serious Games.

References

Abt, C. C. (1970). *Serious games.* Viking Press.

Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthanarayanan, V. (2017). *NMC horizon report: 2017 higher education edition*Austin, Texas: The New Media Consortium.

ADL (2012). *Experience API.* Retrieved March 20, 2016, from https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/.

Alonso-Fernández, C., Caballero Roldán, R., Freire, M., Martinez-ortiz, I., & Fernández-Manjón, B. (2019). *Predicting students' knowledge after playing a serious game based on learning analytics data.* IEEE Access (under review).

Álvarez-García, D., Núñez Pérez, J. C., & Dobarro González, A. (2013). Cuestionarios para evaluar la violencia escolar en educación primaria y en educación secundaria: CUVE3-EP y CUVE3-ESO. *Apuntes de Psicología, 31*(2), 191–202. Retrieved from http://www.apuntesdepsicologia.es/index.php/revista/article/view/322/296.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., et al. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003.

Calvo Morata, A. (2017). *Videojuegos como herramienta educativa en La escuela: Concienciando sobre El ciberbullying*(Master Thesis). Complutense University of Madrid.

Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2018). Making understandable game learning analytics for teachers. *17th international conference on web-based learning* (pp. 112–121). Springer. ICWL 2018) https://doi.org/10.1007/978-3-319-96565-9_11.

Calvo-Morata, A., Rotaru, D. C., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2018). Validation of a cyberbullying serious game using game analytics. *IEEE Transactions on Learning Technologies, 1*, 1. https://doi.org/10.1109/TLT.2018.2879354.

Cano, A. R., Fernández-Manjón, B., & García-Tejedor, Á. J. (2016). Downtown , a subway Adventure : Using learning analytics to improve the development of a learning game for people with intelectual disabilities. *ICALT 2016 - 16th IEEE international conference on advanced learning technologies*https://doi.org/10.1109/ICALT.2016.46.

Cano, A. R., Fernández-Manjón, B., & García-Tejedor, Á. J. (2018). Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. *British Journal of Educational Technology, 49*(4), 659–672. https://doi.org/10.1111/bjet.12632.

Chaudy, Y., Connolly, T., & Hainey, T. (2014). Learning analytics in serious Games : A review of the literature. *Ecaet, 2014* March 2016.

El Asam, A., & Samara, M. (2016). Cyberbullying and the law: A review of psychological and legal challenges. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2016.08.012.

Entertainment Software Association (2017). 2017 essential facts about the computer and video game industry. *Entertainment Software Assotiation, 4*(1), 1–20. Retrieved from http://www.theesa.com/wp-content/uploads/2017/09/EF2017_Design_FinalDigital.pdf%0Ahttp://www.theesa.com/facts/pdfs/ESA_EF_2008.pdf.

European Resuscitation Council Guidelines Writing Group. (2015). *European resuscitation Council guidelines for resuscitation.* Retrieved March 6, 2017, from http://ercguidelines.elsevierresource.com/european-resuscitation-council-guidelines-resuscitation-2015-section-1-executive-summary/fulltext.

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. *Learning, design, and technology* (pp. 1–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_21-1.

Griebel, T. (2006). Self-portrayal in a simulated life: Projecting personality and values in the sims 2. *Game Studies, 6*(1).

Hernández-Lara, A. B., Perera-Lluna, A., & Serradell-López, E. (2019). Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior, 92*, 600–612. https://doi.org/10.1016/j.chb.2018.03.001.

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016). Using game analytics to evaluate puzzle design and level progression in a serious game. *Proceedings of the sixth international conference on learning analytics & knowledge - LAK '16* (pp. 440–448). New York, New York, USA: ACM Press. https://doi.org/10.1145/2883851.2883953.

Jaccard, D., Hulaas, J., & Dumont, A. (2017). In J. Dias, P. A. Santos, & R. C. Veltkamp (Vol. Eds.), *Using comparative behavior analysis to improve the impact of serious games on students' learning experience: Vol. 10653*Cham: Springer International Publishinghttps://doi.org/10.1007/978-3-319-71940-5.

Kato, P. M., & Klerk, S. De (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology, 18*, 1–6.

Kiili, K., Moeller, K., & Ninaus, M. (2018). Evaluating the effectiveness of a game-based rational number training - in-game metrics as learning indicators. *Computers & Education, 120*, 13–28. https://doi.org/10.1016/j.compedu.2018.01.012.

Klimmt, C., Hefner, D., Vorderer, P., Roth, C., & Blake, C. (2010). Identification with video game characters as automatic shift of self-perceptions. *Media Psychology, 13*(4), 323–338. https://doi.org/10.1080/15213269.2010.524911.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073–1137. https://doi.org/10.1037/a0035618.

Kwon, J., & Lee, Y. (2016). Serious games for the job training of persons with developmental disabilities. *Computers & Education, 95*, 328–339. https://doi.org/10.1016/j.compedu.2016.02.001.

Larrañaga, E., Yubero, S., Ovejero, A., & Navarro, R. (2016). Loneliness, parent-child communication and cyberbullying victimization among Spanish youths. *Computers in Human Behavior, 65*, 1–8. https://doi.org/10.1016/j.chb.2016.08.015.

Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning analytics as an assessment tool in serious games: A review of literature. *Serious games and edutainment applications* (pp. 537–563). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-51645-5_24.

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. *Serious games analytics* (pp. 3–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_1.

Long, P., Siemens, G., Gráinne, C., & Gašević, D. (2011). LAK '11 : Proceedings of the 1st international conference on learning analytics and knowledge, february 27 - march 1, 2011, banff, alberta, Canada. *1st international conference on learning analytics and knowledge* (pp. 195). . Retrieved from https://dl.acm.org/citation.cfm?id=2090116.

Marchiori, E. J., Ferrer, G., Fernandez-Manjon, B., Povar-Marco, J., Suberviola, J. F., & Gimenez-Valverde, A. (2012). Video-game instruction in basic life support manuevers. *Emerge, 24*(6), 433–437.

Newman, J. (2002). The myth of the ergodic videogame. *New Media & Society, 4*(3), 405–422. Retrieved from http://www.gamestudies.org/0102/newman/.

Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Evaluar el bullying y el cyberbullying validación española del EBIP-Q y del ECIP-Q. *Psicologia Educativa, 22*(1), 71–79. https://doi.org/10.1016/j.pse.2016.01.004.

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice, 4*(2), 148–169. https://doi.org/10.1177/1541204006286288.

Perez-Colado, I. J., Alonso-Fernández, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2018). Game Learning Analytics is not informagic!. *IEEE global engineering education conference (EDUCON)*.

Pew Research Center (2018). *Teens.* Social Media & Technology 2018.

Sclater, N. (2017). In Routledge (Ed.). *Learning analytics explained.* New York and London: Taylor & Francis Group.

Seif El-Nasr, M., Drachen, A., & Canossa, A. (2013). In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.). *Game analytics*London: Springer Londonhttps://doi.org/10.1007/978-1-4471-4769-5.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious

games. *Computer Standards & Interfaces, 50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014.
Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4.
Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A meta-analysis of data collection in serious games research. *Serious games analytics* (pp. 31–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_2.
Tlili, A., Essalmi, F., Jemni, M., & Kinshuk (2016). An educational game for teaching computer architecture: Evaluation using learning analytics. *2015 5th international conference on information and communication technology and accessibility*ICTA *2015*https://doi.org/10.1109/ICTA.2015.7426881.

6.2. Conference publications

This section contains the publications published as part of conferences or congresses. The following subsections present in detail each publication, full citation, abstract and full text of the publication. As an overview, the conference publications included in the thesis are the following:

1. **Systematizing game learning analytics for serious games:** this publication presents the initial work conducted to systematize GLA for serious games with default analysis and visualizations. The process and results of this work are included in this thesis as part of the results, in subsection 4.5.
2. **Data science meets standardized game learning analytics:** this publication presents the tool T-MON, an exploratory analysis tool of game interaction data, aiming to help in the evidence-based assessment process of players using serious games. The process and results of this work are included in this thesis as part of the results, in subsection 4.4.
3. **Full lifecycle architecture for serious games: integrating game learning analytics and a game authoring tool:** this publication presents some of the earlier work to explore the improvements obtained applying GLA in the serious games' life cycle. The process and results of this work are included in this thesis as part of the results, in subsection 4.5.
4. **Improving serious games analyzing learning analytics data: lessons learned:** this publication presents some of the earlier work to explore the improvements obtained applying GLA in the serious games' life cycle. The process and results of this work are included in this thesis as part of the results, in subsection 4.5.
5. **Applications of learning analytics to assess serious games:** this publication presents some of the earlier work to explore the opportunities for assessment using GLA data with serious games. The process and results of this work are included in this thesis as part of the results, in subsection 4.5.

6.2.1. Systematizing game learning analytics for serious games

Full citation

Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2017): **Systematizing game learning analytics for serious games.** IEEE Global Engineering Education Conference (EDUCON), 25-28 April 2017, Athens, Greece.

This paper received a **Best Paper Award** of the Conference, in the "Area 3: Innovative Materials, Teaching and Learning Experiences in Engineering Education".

Abstract

Applying games in education provides multiple benefits clearly visible in entertainment games: their engaging, goal-oriented nature encourages students to improve while they play. Educational games, also known as Serious Games (SGs) are video games designed with a main purpose other than pure entertainment; their main purpose may be to teach, to change an attitude or behavior, or to create awareness of a certain issue. As educators and game developers, the validity and effectiveness of these games towards their defined educational purposes needs to be both measurable and measured. Fortunately, the highly interactive nature of games makes the application of Learning Analytics (LA) perfect to capture students' interaction data with the purpose of better understanding or improving the learning process. However, there is a lack of widely adopted standards to communicate information between games and their tracking modules. Game Learning Analytics (GLA) combines the educational goals of LA with technologies that are commonplace in Game Analytics (GA), and also suffers from a lack of standards adoption that would facilitate its use across different SGs. In this paper, we describe two key steps towards the systematization of GLA: 1), the use of a newly-proposed standard tracking model to exchange information between the SG and the analytics platform, allowing reusable tracker components to be developed for each game engine or development platform; and 2), the use of standardized analysis and visualization assets to provide general but useful information for any SG that sends its data in the aforementioned format. These analysis and visualizations can be further customized and adapted for particular games when needed. We examine the use of this complete standard model in the GLA system currently under development for use in two EU H2020 SG projects.

Full publication

Systematizing game learning analytics for serious games

Cristina Alonso-Fernandez, Antonio Calvo, Manuel Freire, Ivan Martinez-Ortiz, Baltasar Fernandez-Manjon
Dept. Software Engineering and Artificial Intelligence
Universidad Complutense de Madrid, Facultad de Informática
C/ Profesor JoseGarciaSantesmases, 9 28040 Madrid, Spain
{crisal03, antcal01}@ucm.es {manuel.freire, imartinez, balta}@fdi.ucm.es

Abstract **– Applying games in education provides multiple benefits clearly visible in entertainment games: their engaging, goal-oriented nature encourages students to improve while they play. Educational games, also known as Serious Games (SGs) are video games designed with a main purpose other than pure entertainment; their main purpose may be to teach, to change an attitude or behavior, or to create awareness of a certain issue. As educators and game developers, the validity and effectiveness of these games towards their defined educational purposes needs to be both measurable and measured. Fortunately, the highly interactive nature of games makes the application of Learning Analytics (LA) perfect to capture students' interaction data with the purpose of better understanding or improving the learning process. However, there is a lack of widely adopted standards to communicate information between games and their tracking modules. Game Learning Analytics (GLA) combines the educational goals of LA with technologies that are commonplace in Game Analytics (GA), and also suffers from a lack of standards adoption that would facilitate its use across different SGs. In this paper, we describe two key steps towards the systematization of GLA: 1), the use of a newly-proposed standard tracking model to exchange information between the SG and the analytics platform, allowing reusable tracker components to be developed for each game engine or development platform; and 2), the use of standardized analysis and visualization assets to provide general but useful information for any SG that sends its data in the aforementioned format. These analysis and visualizations can be further customized and adapted for particular games when needed. We examine the use of this complete standard model in the GLA system currently under development for use in two EU H2020 SG projects.**

Keywords—game analytics; serious games; e-learning; dashboard; xAPI

I. Introduction

The success of games for entertainment purposes, especially among younger generations, has made them of interest for researchers in different fields such as mathematics, physics, engineering, medicine, economics, history or literature [1]-[2]-[3]-[4]. In the field of education, their engaging and goal-oriented nature encourages students to outdo themselves while learning key concepts derived from educational plans.

Educational games, also known as Serious Games (SGs) are video games designed not only for pure entertainment; their main purpose may be to teach, to change an attitude or behavior or to create awareness of a certain issue. Throughout the years, many serious games have had great success in achieving their different educational purposes (e.g. teaching Mathematics, English, social abilities or change an attitude towards certain problems). For instance, the serious game *Darfur is Dying* was launched in April of 2006 to help to shed a light on the ongoing war in the Darfur region of Sudan at that time and the humanitarian disaster derived from it for 2.5 million refugees. Despite the uncertainty on the consequences it may have brought to the actual crisis, the game attracted 800.000 players in only 6 months [5]. *Foldit* is an online puzzle serious game on protein folding that helped decipher the crystal structure of the M-PMV retroviral protease, of importance to antiretroviral drug development. With the help of thousands of players competing against each other, an accurate model of the enzyme was found in only 10 days, while the answer had troubled medical science for the preceding 15 years [6]. The adventure serious game *Aislados*, which aims to teach abilities that help to prevent drug addiction, sexist behavior and other risk behaviors among teenagers, has received several awards for its help in attitude change [7]. *Treefrog Treasure* is a platformer serious game to teach players fractions, rational numbers and percentages. While collecting in-game jewels, players control the frog's jumps through barriers that contain mathematical questions, teaching players the placement of fractions and rational numbers on the number-line that the barrier represents. When players make a mistake, hints are provided to help them to find the correct answer [8].

When applying games in education, providing access to information on the interactions of students with the game is not only desirable, but essential. Proving their validity and effectiveness is integral to their educational purpose and to provide the means to evaluate the knowledge obtained by the students through their game-play [9].

A common method to formally evaluate SGs consists of carrying out a test with the students before and after playing the game, and comparing the results through statistical analysis [10]. This pre-post method is both expensive and time consuming, and provides very limited information regarding the student's learning process. As a consequence, very few games have been formally proved to be effective. Basic information from students has also been collected with learning management systems (LMS) providing a brief insight into student actions [9] but still failing to explain how students learn.

However, as games are highly interactive digital content, a different approach can be used. In e-learning, it is common to

978-1-5090-5467-1/17/$31.00 ©2017 IEEE 25-28 April 2017, Athens, Greece
2017 IEEE Global Engineering Education Conference (EDUCON)

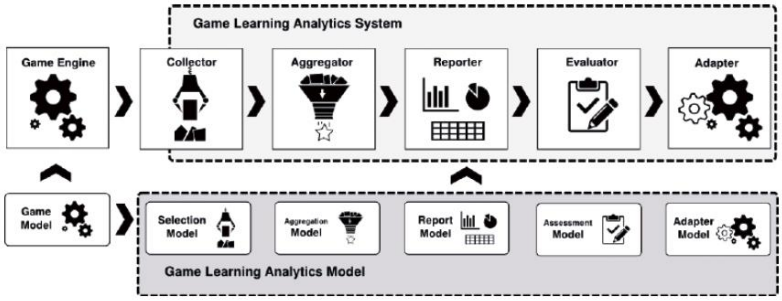

Figure 1. Game Learning Analytics (GLA) conceptual architecture model. The game sends data to a collector for its aggregation. The information obtained is used to feed reports, visualizations, evaluate students and, through the adapter, to turn into instructions that go back to the game. [9]

use Learning Analytics (LA) to capture interaction data with the purpose of better understanding or improving the learning process. LA can be defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [11]-[12]. To apply this to serious games, interaction data can be captured by adding a tracker to the SG that sends player interaction data (also referred to as traces) to a server. Analysis of the traces can yield actionable information regarding the students' interactions with the game, making the set of actions, mistakes and correct actions of the player meaningful [13].

Without analytics, SGs in education are akin to black boxes: they merely provide a final state that shows the students' gameplay results, usually in the form of simple metrics such as the player's final score; but that does not provide information regarding the learning process. Opening the box can provide much more information on the use of SGs and how their players interact with them. For instance, through aggregation of data, game developers can determine which game areas present greater challenges for players, which ones are easier and which ones can be improved, either due to excessive complexity, because they suffer from some design problem or because players do not understand what they are expected to do. Eventually, if the game is well designed and the relevant interaction data is captured, it should be possible to trace the evolution of each player's knowledge at every part of the game, and identify any areas where they struggle or shine.

II. GAME LEARNING ANALYTICS

To obtain the desired information from games, we propose the following complete, scalable, standards-based analytics architecture. We describe the architecture in greater detail in following sections.

A. Educational scenarios

Within this architecture, games send traces to a server that analyzes the data and transforms it into useful information which is then displayed and explored by stakeholders: the teacher or instructor in charge of players, the players themselves, the game developer or designer, and the researchers. This information can be used in multiple ways:

- At run-time by the game itself, which can then adapt its characteristics on the fly, providing a personalized, adaptive experience.
- At run-time by teachers and instructors, which can use it to locate students that are experiencing problems and help these students out.
- After a game session, to understand how the game was played and measure student knowledge acquisition, allowing the game to be revised and improved for future players; and providing valuable feedback to players regarding the session.
- After a game session, to evaluate students based on their performance. Note that this is the only use-case where pre-test / post-test evaluation would also work.

B. Learning analytics and Game analytics

One of the main problems of LA is the lack of widely adopted standards to communicate trace information between games and their tracking modules, due to the ad-hoc nature of different data analysis solutions. As a result, each SG ends up being tied to its own LA solution; whenever the game is updated, game-specific tracking, analysis and visualization assets must be updated in the LA, increasing development costs. As long as such tight coupling between game and analytics is required, LA for SGs – Game Learning Analytics (GLA) – will continue to be rarely used.

Figure 1 shows an abstract overview of a GLA system. The game engine sends data to a collector via its tracking component. The data collected is then aggregated and analyzed, with the results used to feed reports (either in real-time or for later use). This information may also have other purposes, such as supporting assessment. Finally, an (optional) adapter sends instructions derived from that information back to the game [9].

Apart from the educational goals of LA, GLA also feeds off the tools and technologies from Game Analytics (GA). For many years, the industry of entertainment videogames has used GA

978-1-5090-5467-1/17/$31.00 ©2017 IEEE 25-28 April 2017, Athens, Greece
2017 IEEE Global Engineering Education Conference (EDUCON)

extensively to obtain information from their players, also collected through non-disruptive tracking tools embedded in the games [14]. The key difference between GA and LA and GLA is GA's exclusive focus on the game itself. In traditional GA, analytics are only intended for game-developers or, at the very most, to obtain financial information (in the case of games with built-in transactions). Furthermore, traditional GA has no concept of tracking learning, and cannot accommodate teachers that want to explore what players have learnt and, possibly, share results with other teachers for comparison or research purposes.

C. Systematizing Game Learning Analytics

The process of GLA for serious games still suffers from lack of widespread standards. In this paper, we describe two key steps towards the systematization of GLA, and examine their use in the GLA system currently under development for use in two EU H2020 SG projects. The two key contributions are:

- The use of a newly-proposed standard to exchange SG traces between the game and the analytics platform. This allows standard, reusable tracker components to be developed for each game engine or development platform, communicating via this standard with the server-based analytics platform. Furthermore, this allows traces to be shared and analyzed by any tool that can handle this upcoming standard, instead of being limited to the analytics system where they were first captured.
- The use of standardized, modular analysis and visualization assets, which can be customized and adapted for particular games if needed, but that already provide useful information out-of-the-box for any SG that sends its data in the aforementioned format.

Combining both steps together, we can provide a complete GLA system to be applied to any SG, as long as this SG makes use of a compliant tracking component as described in the following section. Without further requirements or configuration, the proposed system can generate meaningful dashboards for different stakeholders (e.g. developers, teachers).

III. Data Tracking

To systematize the tracking step, we use a general tracking model together with its implementation using the Experience API standard. The tracking model was defined after an analysis of the current state of data standards and serious games, in addition to previous experiences applying e-learning standards to serious games [15]-[16]. The resulting interaction model is described in [17]-[18].

A. Experience API (xAPI)

The Experience API (xAPI) is a data format developed by the Advanced Distributed Learning Initiative (ADL) Initiative [19] together with an open community. The xAPI standard derives from Activity Streams, a format that can be used to describe streams composed of actors performing with actions in a specific context. In this sense, each xAPI statement represents a learning activity and has three main attributes: an actor, a verb and an object: *who* did *what* action, with a *target* of the action and certain additional attributes (for example, to provide more context or describe the results of the action), as shown in Figure 2.

```
{
  "actor": {
    "name": "John Doe",
    "mbox": "mailto:johndoe@example.com"
  },
  "verb": {
    "id": "http://adlnet.gov/expapi/verbs/initialized",
    "display": { "en-US": "initialized" }
  },
  "object": {
    "id": "http://rage.e-ucm.com/activities/Countrix",
    "definition": {
      "name": { "en-US": "Countrix Serious Game" },
      "type:" "https://w3id.org/xapi/seriousgames/activities/serious-game
    }
  }
}
```

Figure 2. An example xAPI statement representing the learning activity "John Doe (actor) initialized (verb) the activity serious game Countrix (object)".

The interaction model makes use of concepts such as *completables* (tasks, quests or mini-games with a beginning and an end), *alternatives* and general *variables* to track interactions in the specific domain of SGs. Custom interactions may also be defined to provide further information in a specific SG. The interactions, events and targets are mapped by the tracker library into their corresponding xAPI statement attributes, verbs and activity types respectively. This interaction model and its implementation in the xAPI standard provide a general, game-independent trace format that can model most, and frequently all, the interactions that a player makes with a SG.

We validated the tracking model with a serious game developed specifically for this purpose: the geography Q&A game Countrix [20]. The game consists of a series of multiple-choice questions (capital, country, continent or flag). Each correct answer increases the score, while each wrong answer decreases the remaining time to play; the game can display the xAPI statements generated either during game-play or when the player runs out of time.

B. Key performance indicators

Obtaining useful information from gameplays may require dealing with large amounts of data, sent by connected clients at high rates. For example, a game that sends many traces and is being played simultaneously by many students could easily overload a naïve collector implementation. However, scalability by itself is not enough: trace data must be analyzed to be useful for stakeholders, and there is considerable value in performing the analysis while the game is being played, instead of only once the session finished. For example, near real-time data allows teachers to perform interventions on students that are still playing, and, should the game contain adaptive mechanics, allows the game to adapt itself to the player's actions.

Not all classical gamification metrics are suitable or useful for learning and training. For instance, it is common to compare a student's results with the average of the class. Although this metric may provide an idea of how well the student is doing in the course, it could also discourage those students whose score falls below the class average. Speed, in terms of actions per time unit, is another commonly measured metric in entertainment games that may lead students to rush, being negatively

correlated with performance [13]. To provide suitable metrics, educators using our system can provide quantifiable outcomes as key performance indicators (KPIs) to be measured in the analytics. KPIs are used to measure players' performance denoting their level of success, usually through a quantitative indicator. Examples of KPIs may be the number of errors made or the percentage of game completion.

C. *Issues when collecting data*

All xAPI statements generated by the game will be collected in a Learning Record Store (LRS). The LRS concept derives from the e-learning domain as a database system to store statements in sequential order [21]. Supporting xAPI input and output, an LRS typically allows authenticated and authorized users to save and query traces.

Collecting data from players' interactions requires awareness of applicable personal privacy laws and regulations. This issue becomes particularly important in domains such as education with underage students or when dealing with health-related data. An important part of complying with these laws cannot be performed from within a GLA tool: for example, players should be provided with informed-consent forms before collecting any data; and these forms should clearly state information such as the ownership of the collected data, how the data will be used, for which purposes, and who can use the data under which circumstances [9].

Anonymization is another key issue when collecting data. Personal information should either not be collected at all, or it should be anonymized immediately after collection [9]. Users of our GLA system are limited both as to the data they can access and the level of anonymization with which it is provided. Teachers have full access to data from their students; but game developers and researchers can only access anonymized and/or aggregated data.

Finally, when experiments produce interesting data, allowing it to be reused by researchers from different domains with different purposes can greatly increase its value. For example, the OpenAIRE2020 Project seeks to create an open infrastructure for research in Europe, sharing open data (i.e. data free to access, reuse, repurpose and redistribute) in open repositories to make it available for other researchers [22]. xAPI for SGs provides a good candidate for sharing SG-related activity data.

IV. Data Analysis and Visualization

To systematize the analysis and visualization steps, there are two design goals to be met:

1) Given no knowledge of the game beyond the xAPI traces that it produces, which are assumed to comply with the xAPI SG recipe, we are interested in a default set of analyses and visualizations that provides as much insight as possible for zero customization cost.
2) Advanced users must be allowed to add game-specific information to create tailored dashboards and visualizations – while minimizing the amount of configuration, and allowing the resulting visualizations to be reused between SGs with similar requirements.

The analysis performed on the tracked data should focus on the suitable metrics and KPIs defined by educators, avoiding metrics that may confuse students or work against their learning process. Additionally, the default set of analysis of visualizations should also be adapted to the needs and interests of the different stakeholders involved in the process of GLA: teachers, students, game developers or designers, managers and researchers. With these considerations, we propose the set of default visualizations depicted in Table 1, each of which is geared towards a specific stakeholder and may require a specific underlying analysis. Note that certain visualizations may be of interest to multiple stakeholders; Table 1 only lists the primary stakeholder.

TABLE I. Default Set of Visualizations for Stakeholders

Visualization description	**Primary Stakeholder**
For all students in the class, or individually selected students: sessions, questions answered, total errors, ratio of correct answers, timestamps	Developer
Distribution of scores obtained in the game	
Distribution of questions answered by students	
Times each question has been answered	
Number of times each *accessible* has been accessed per player	
Duration of sessions	
Number of times each Experience API (xAPI) verb has been used	
Use of xAPI verbs over time	
Peak times of game use	Manager
Inter-group comparisons	
For the student: questions answered, errors, ratio of correct answers, final score, timestamps, session duration	Student
Users ranked by number of errors	Teacher
Questions with higher error ratio	
Number of players in each game-play session	
Total number of correct and incorrect alternatives selected in multiple-choice questions by each player	
Total number of correct and incorrect alternatives selected by players in each multiple-choice question	
Alternatives selected in each multiple-choice question	
Score achieved by players in the different completables	
Progress achieved by players in the different completables	
Progress of players over time	
For each video in the game, the number of times it has been seen and skipped by players	

978-1-5090-5467-1/17/$31.00 ©2017 IEEE
25-28 April 2017, Athens, Greece
2017 IEEE Global Engineering Education Conference (EDUCON)

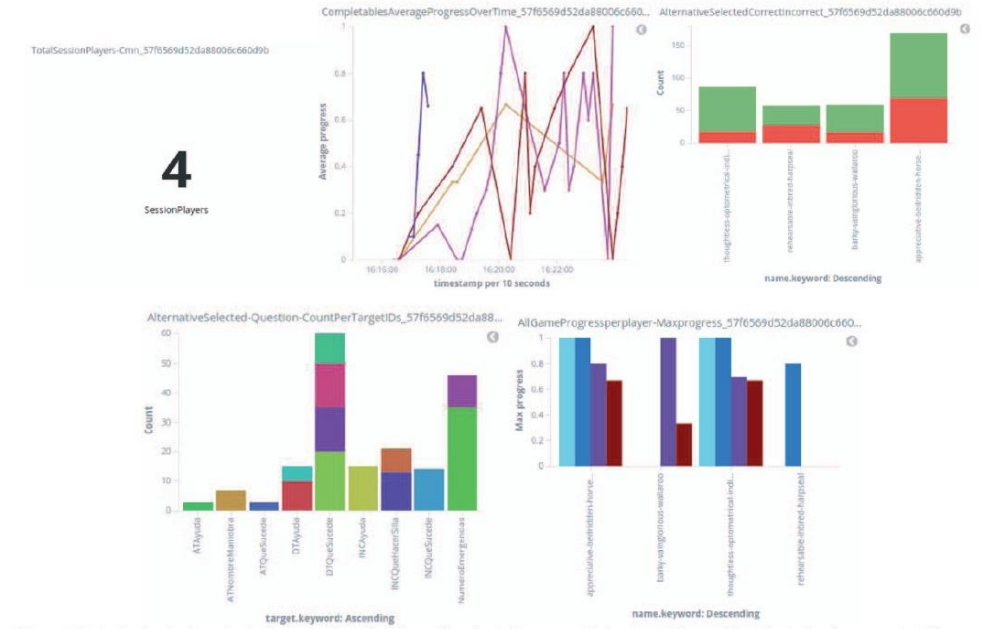

Figure 3. Default visualizations for teachers: number of sessions; line chart of progress of players per time and bar charts showing correct and incorrect alternatives selected in multiple-choice questions, the alternatives selected and progress achieved in the different completables of the game.

Figure 4. Some of the default visualizations for developers: pie chart showing the number of times each accessible has been accessed per player; line chart showing the use of xAPI verbs over time; bar chart showing the number of times each xAPI verb has been used; and bar chart showing for each video the number of times it has seen and skipped by players.

A. Analysis and visualization tools

The information obtained from the analysis is stored in Elasticsearch [23], which can analyze and search a vast amount of data in near real time (with delays measured in few seconds). The visualization dashboards have been developed with Kibana [24], an open source platform that provides a flexible browser-based interface to quickly develop analysis and visualizations. Once created, these dashboards can change dynamically to display updated results as they become available. Our GLA system currently supports all visualizations listed in Table 1.

Figure 3 displays some of the default visualizations for teachers, including total number of sessions registered, progress of players per time; correct and incorrect alternatives selected and progress achieved in multiple-choice questions.

Figure 4 shows some of the relevant information for developers including number of times each accessible has been accessed and use of xAPI verbs over time.

B. Personalization of analysis and visualizations

To allow these analyses and visualizations to be configurable and reusable, we have developed a simple wizard-based tool that allows users to first choose the desired visualization templates, and later connect these templates to the chosen analysis outputs.

For teachers, these dashboards provide both overviews and details of students' interactions with the game. In particular, teachers can easily switch between individual student view and whole-class view. Many of the whole-class visualizations have direct interpretations in terms of educational interventions. For example, from a visualization showing the number of errors made by students in multiple-choice questions, teachers can quickly zero in on those with higher error ratio for review with their students.

C. Alerts and warnings

Since analyses are near real time, we have included the possibility of configuring alerts and warnings, mostly intended for teachers that may be present during a SG session. A warning is a message displayed when a certain condition is satisfied (for example, "a student has been inactive for two minutes"); alerts are similar to warnings, but are intended for situations that require immediate action from the teacher, such as "a student has answered a very important question wrong". During a session, users with outstanding alerts and warnings are marked with icons. When clicking on any user, the full descriptions of any applicable alerts and warnings will be displayed (see Figure 5).

We have only identified a single generic alert ("a student has been inactive for 2 minutes") for default inclusion. In general, unlike visualizations, alerts and warnings are highly game-specific.

D. Data exploitation

The information obtained through the visualizations can be used for several purposes:

- Students' assessment: analyzing the statements obtained from the students' gameplay, teachers can evaluate students. From the learning plan, educational goals will become tasks or levels that students need to complete to pass a certain topic in the course evaluation. Scores could indicate a mark for students.
- Personalized and adaptive gameplays: if correctly tracked and analyzed in near real-time, the collected data could be used to personalize the gameplay while students are still playing. Through that adaptive learning experience, students could benefit from the interaction with the game.
- Serious games' improvement: the results will also provide an insight into the serious game's accuracy and suitability. Game bugs or errors, unreachable areas, places with lack of information or difficult to understand and tasks too difficult for students are some of the information obtained from analyzing statements.

Figure 5. Alerts (messages that appear when a certain condition is satisfied and required immediate action from the teacher) and warnings (also result from a true condition but do not require immediate action) can be configured in the architecture so teachers can receive those notifications in the real-time view. For each user, the number of alerts and warnings satisfied as a result of the user's gameplay appear. User-specific information can be obtained by clicking on a single user, including the alerts or warnings' descriptions that user's gameplay has given rise to.

V. GLA Architecture

Providing a full GLA system requires handling multiple tasks, from data acquisition via tracking to collection, data analysis, and result visualization. A diagram of the proposed GLA architecture, which comprises several modules that work together to analyze data and visualize results, can be found in Figure 6.

- The game design, learning goals and learning design determine the design and implementation of the game (its mechanics, goals and characters). Both the learning and game design are essential as they determine the elements (usually variables) that will appear in the game design containing educational information. Those elements are the ones that should be tracked, as

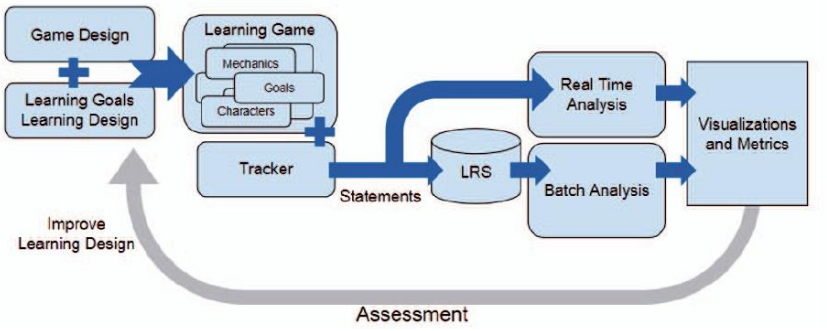

Figure 6. Overview of Game Learning Analytics (GLA): from learning goals, learning design and game design; the tracker embedded in the game sends Experience API (xAPI) statements to a Learning Record Store (LRS) for batch analysis, and directly to the real-time analysis. Some visualizations and metrics may be derived from the analysis to obtain further information for students' assessment and learning design improvement.

they can reveal whether players are actually learning or not.

- The game uses a generic *Tracker* component to send the standardized xAPI statements. The tracker components are available in Unity, Java and C#, for their easy integration with different serious games.
- After passing through an authorization and authentication module, the statements are saved into a Learning Record Store (LRS) that supports xAPI format, and also submitted to a real-time analysis component which calculates an updated state-of-game for each player. Should a different analysis be desired, the LRS can be used to replay the relevant statements; this is labeled as "batch analysis".
- Finally, the data is displayed in suitable dashboards comprising all relevant metrics for use by the relevant stakeholders. Personalized dashboards as well as configured alerts and warnings are also displayed.
- The process ends when the information obtained is reintroduced as improvements in the game learning design or as personalized and adaptive gameplays for students. Finally, the information may also help in the students' assessment process.

All components have been developed as open source and are available online[1].

VI. Conclusions and Future Work

Game Learning Analytics (GLA) for serious games is no longer an emerging field; however, it is still performed mostly through ad-hoc analysis, and can greatly benefit from a more systematic, standardized approach. Doing so would make GLA easier to apply to new serious games, greatly reducing the costs of building ad-hoc tools, and helping to drive teacher adoption of serious games. Teacher involvement requires GLA systems that provide value out of the box, and with minimal configuration; while still allowing advanced users to customize and tailor their analytics to specific requirements. We have proposed two important steps to achieve this systematization: First, the use of a an Experience API (xAPI) recipe specific for serious games to allow standardized trace collection and enable sharing; and second, a set of visualizations that we consider to be useful for a wide variety of possible SGs. We have also described the architecture of a system that uses both of these steps. This system is available as open software, and we plan to extend and validate it in multiple educational scenarios during the following months, as part of two EU H2020 SG-related projects.

Currently, creating a template requires using Kibana's built-in visualization authoring environment; but we are developing a wrapper around this that should make the process simpler for non-programmers. We are also currently working on the integration of the GLA architecture with LTI (Learning Tools Interoperability) [25] and SAML2 (Security Assertion Markup Language v 2.0) [26] to manage authentication. These would greatly decrease the amount of configuration needed to register students and teachers for large-scale deployments in educational institutions that already rely on one of these technologies.

Acknowledgment

This work has been partially funded by Regional Government of Madrid (eMadrid S2013/ICE-2715), by the Ministry of Education (TIN2013-46149-C2-1-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676).

[1] eUCM Research Group, RAGE Analytics, (2016). https://github.com/e-ucm/rage-analytics

REFERENCES

[1] M.J. Mayo, Video Games: A Route to Large-Scale STEM Education?, Science (80-.). 323 (2009) 79–82. doi:10.1126/science.1166900.

[2] B.M. Iglesias, C. Fernandez-Vara, B. Fernandez-Manjon, E-Learning Takes the Stage: From La Dama Boba to a Serious Game, IEEE Rev. Iberoam. Tecnol. Del Aprendiz. 8 (2013) 197–204. doi:10.1109/RITA.2013.2285023.

[3] K.H. Evans, W. Daines, J. Tsui, M. Strehlow, P. Maggio, L. Shieh, Septris: a novel, mobile, online, simulation game that improves sepsis recognition and management., Acad. Med. 90 (2015) 180–4.

[4] K. Squire, M. Barnett, J.M. Grant, T. Higginbotham, Electromagnetism supercharged!:learning physics with digital simulation games, Int. Conf. Learn. Sci. (2004) 513–520.

[5] L. interFUEL, Darfur is Dying, (2006). http://www.gamesforchange.org/play/darfur-is-dying/.

[6] Center for Game Science at University of Washington in collaboration with UW Department of Biochemistry., Foldit: Solve Puzzles for Science, (2008). http://fold.it/portal/ (accessed November 4, 2016).

[7] Asociación Servicio Interdisciplinar de Atención a las Drogodependencias (SIAD), Aislados, (2014). http://www.aislados.es/zona-educadores/ (accessed November 13, 2016).

[8] Center for Game Science at the University of Washington, Treefrog Treasure, (2016). http://centerforgamescience.org/blog/portfolio/treefrog-treasure/ (accessed November 15, 2016).

[9] M. Freire, Á. Serrano-Laguna, B.M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, B. Fernández-Manjón, Game Learning Analytics: Learning Analytics for Serious Games, in: Learn. Des. Technol., Springer International Publishing, Cham, 2016: pp. 1–29. doi:10.1007/978-3-319-17727-4_21-1.

[10] A.C. and M. Ruiz, A systematic literature review on serious games evaluation: An application to software project management, Comput. Educ. 87 (2015) 396–422.

[11] G. Long, P., & Siemens, Penetrating the Fog: Analytics in Learning and Education, Educ. Rev. (2011) 31–40.

[12] G. Siemens, G., Dawson, S., & Lynch, Improving the Quality and Productivity of the Higher Education Sector. Policy and Strategy for Systems-Level Deployment of Learning Analytics., (2013).

[13] C.S. Loh, Y. Sheng, D. Ifenthaler, Serious Games Analytics, Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-05834-4.

[14] S.S. and M. Vaden, Telemetry and Analytics Best Practices and Lessons Learned, Game Anal. Maximizing Value Play. Data. (2013) 85–109.

[15] A. Serrano, E.J. Marchiori, A. del Blanco, J. Torrente, B. Fernandez-Manjon, A framework to improve evaluation in educational games, in: Proc. 2012 IEEE Glob. Eng. Educ. Conf., IEEE, 2012: pp. 1–8. doi:10.1109/EDUCON.2012.6201154.

[16] Á. del Blanco, E.J. Marchiori, J. Torrente, I. Martínez-Ortiz, B. Fernández-Manjón, Using e-learning standards in educational video games, Comput. Stand. Interfaces. 36 (2013) 178–187. doi:10.1016/j.csi.2013.06.002.

[17] Á. Serrano-Laguna, I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, B. Fernández-Manjón, Applying standards to systematize learning analytics in serious games, Comput. Stand. Interfaces. (2017) 116–123. doi:http://dx.doi.org/10.1016/j.csi.2016.09.014.

[18] eUCM Research Group, xAPI Serious Games Profile, (2016). http://w3id.org/xapi/seriousgames (accessed October 27, 2016).

[19] ADL Initiative, xAPI Specification, 2014. (2016). https://github.com/adlnet/xAPI-Spec/blob/a752217060b83a2e15dfab69f8c257cd86a888e6/xAPI.md (accessed October 27, 2016).

[20] eUCM Research Group, Countrix Serious Game, (2016). https://github.com/e-ucm/countrix (accessed October 27, 2016).

[21] Scorm, what is an lrs learning record store, (2016). http://scorm.com/tincanoverview/what-is-an-lrs-learning-record-store/ (accessed November 4, 2016).

[22] H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, (2016). http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (accessed October 31, 2016).

[23] Elastic, Elasticsearch, (n.d.). https://www.elastic.co/products/elasticsearch (accessed March 20, 2016).

[24] Elastic, Kibana, (n.d.). https://www.elastic.co/products/kibana (accessed March 18, 2016).

[25] IMS Global Learning Consortium, Learning Tools Interoperability, (2016). https://www.imsglobal.org/activity/learning-tools-interoperability (accessed November 13, 2016).

[26] OASIS, Security Assertion Markup Language (SAML), (2005). https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security (accessed November 13, 2016).

978-1-5090-5467-1/17/$31.00 ©2017 IEEE 25-28 April 2017, Athens, Greece
2017 IEEE Global Engineering Education Conference (EDUCON)

6.2.2. Data science meets standardized game learning analytics

Full citation

Cristina Alonso-Fernández, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2021): **Data science meets standardized game learning analytics.** IEEE Global Engineering Education Conference (EDUCON), 21-23 April 2021, Vienna, Austria.

Abstract

Data science applications in education are quickly proliferating, partially due to the use of LMSs and MOOCs. However, the application of data science techniques in the validation and deployment of serious games is still scarce. Among other reasons, obtaining and communicating useful information from the varied interaction data captured from serious games requires specific data analysis and visualization techniques that are out of reach of most non-experts. To mitigate this lack of application of data science techniques in the field of serious games, we present T-Mon, a monitor of traces for the xAPI-SG standard. T-Mon offers a default set of analysis and visualizations for serious game interaction data that follows this standard, with no other configuration required. The information reported by T-Mon provides an overview of the game interaction data collected, bringing analysis and visualizations closer to non-experts and simplifying the application of serious games.

Full publication

Data science meets standardized game learning analytics

Cristina Alonso-Fernández
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
calonsofernandez@ucm.es

Antonio Calvo-Morata
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
acmorata@ucm.es

Manuel Freire
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
manuel.freire@fdi.ucm.es

Iván Martínez-Ortiz
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
imartinez@fdi.ucm.es

Baltasar Fernández Manjón
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
balta@fdi.ucm.es

***Abstract*—Data science applications in education are quickly proliferating, partially due to the use of LMSs and MOOCs. However, the application of data science techniques in the validation and deployment of serious games is still scarce. Among other reasons, obtaining and communicating useful information from the varied interaction data captured from serious games requires specific data analysis and visualization techniques that are out of reach of most non-experts. To mitigate this lack of application of data science techniques in the field of serious games, we present T-Mon, a monitor of traces for the xAPI-SG standard. T-Mon offers a default set of analysis and visualizations for serious game interaction data that follows this standard, with no other configuration required. The information reported by T-Mon provides an overview of the game interaction data collected, bringing analysis and visualizations closer to non-experts and simplifying the application of serious games.**

Keywords—serious games, learning analytics, xAPI, dashboards, data science, visual analytics

I. Introduction

Serious Games (SGs) is a broad term that encompasses any game with a main purpose beyond entertainment [1]. Typical purposes include teaching knowledge, raising awareness about issues, or changing the attitudes or behaviors of its players. SGs have been applied in a wide range of fields, including education, healthcare, communication, or politics [2]. In the educational field, COTS (Commercial Off-the-Shelf Games) videogames can also be used for educational purposes [3], although there may be more barriers to adopt them in the classroom compared to serious games.

The application of games in educational scenarios presents multiple benefits: games provide an immersive learning environment, where risky or complex scenarios can be tested in safety while providing immediate feedback to players about their actions, and breaking the common 10-minute barrier of attention [4]. In this way, videogames allow the player to play an active role in their learning process.

Despite these benefits, the application of serious games is still limited. Among the barriers that exist when applying the videogame in the classroom are: the limited duration of typical class periods vs. that of games, the lack of definition of the role of teachers during game sessions, together with their low familiarity with serious games; the hardware infrastructure of schools; and the lack of resources to evaluate and track student progress [5]. Among these limitations, we highlight the fact that educators do not have information about what is happening in the game; instead, games act as a black box, and teachers have no control or insight into what is happening while students play. Therefore, it becomes very difficult to use this type of learning tool to effectively assess players.

A commonly used technique evaluate players is to make them fill out a questionnaire before playing the game, and a subsequent questionnaire after playing the game, and then compare both responses to measure the effect of the game on its players [6]. This methodology, however, also has drawbacks, as the measurement of learning is carried out externally, outside the learning environment, and taking a questionnaire could have additionally negative effects on players' performance [7]. Moreover, when only applying questionnaires, educators do not receive any information about the behavior and choices/answers made by the players, neither during nor after the game.

The problem of user tracking and evaluation is an inherent limitation to the use of new technologies in the classroom. However, it is possible to apply Learning Analytics techniques to address it. Learning Analytics (LA) are defined as: the "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [8]. The educational contexts in which learning analytics are most popular are learning management systems (LMSs) such as Moodle, and in massive open online courses (MOOCs). These massive courses have been the biggest driver of analytics due to the large number of students involved, making it infeasible to track each student individually using traditional methods. This data-based approach also aligns with the current trend in many other fields where data science applications are growing dramatically, and in particular in education, where data-based approaches have been identified as one of the biggest challenges, through the fields of Educational Data Mining and Learning Analytics [9].

In the context of serious games, the application of this technique is called Game Learning Analytics (GLA) [10], and helps to relate gameplay with learning, providing a more evidence-based measure of players' performance. The information gathered from players' interactions can help, first, to validate the game and its design [11], and also to assess

players, as defined in the field of *stealth assessment* [12]. However, the use of these techniques and data science associated with games is very limited due to infrastructure problems in the school, their complexity for educators, and the cost personalizing analyses to each specific serious game.

In this paper, we present T-Mon, a platform to tackle some of the previously identified issues when analyzing interaction data from serious games, by providing a default set of visualizations for any serious game data collected using a standard and validated format. The rest of the paper is structured as follows: Section 2 presents some related work on Game Learning Analytics, and the standard data format used to collect data from serious games (xAPI-SG); Section 3 presents T-Mon, the platform created to simplify the analysis and visualization of data collected from serious games; finally, Section 4 discusses the platform and presents the conclusions of our work.

II. Related Work

Learning Analytics (LA) has the potential to provide precise and evidence-based information about the process and progress of learners in an educational environment. LA has been used for many purposes, such as: enhancing the learning experience, analyzing the impact of interactions between students in learning, supporting the evaluation of learning designs, predicting students at risk of failing, predicting dropout in MOOCs, making sense of multimodal data and, more broadly, modelling players and predicting their performance [13]. However, there are still many remaining challenges, including: compliance with privacy requirements, data heterogeneity and ownership, lack of technology frameworks, and the lack of generalization of applications and tools [14]. Authors have also pointed out the need for evidence of the long-term impact of LA practices on learning and teaching practice.

LA techniques can also be applied to serious games, where players/learners interact with the learning environment (in this case, the game) creating a rich interaction data that can be analyzed for multiple purposes. In particular, the application of LA in the context of serious games has had two main focuses: predicting players performance, and visualizing players results [15]. The large amount of LA data gathered from serious games can also be analyzed with more complex data science techniques to obtain deeper information. Research in this area has focused on predicting the effect of the game on players based on their interactions and creating different players profiles to analyze and understand their learning process in the game [16].

The analysis of interaction data from serious games can be performed both at near real-time (while students are playing) and/or after the gameplays have finished:

- In real-time, the data collected and analyzed can tackle the issue of teachers losing track of players during the application of games: while students play, teachers can receive information about students' actions and progress, gaining insights about their learning and intervening if necessary. These real-time metrics can also be used to evaluate players at real time, comparing results and choices/answers among different students.
- In batch (offline), once all students have finished, more complex analysis techniques on the aggregated data can provide further insights about the results. For instance, players could be clustered or classified, to provide information about the different players' profiles and their learning status and needs. The aggregated results could also be combined and compared with results from other activities, or even with other data sources.

While the use of LA techniques can be effective in many educational aspects related to serious games, its integration and application in real scenarios is complex, from the infrastructure and format of the data to the type of analysis and the goals to be addressed. Chatti et al. presents a model for the application of LA with four dimensions (Fig. 1): the "what" defines the data collected by the system, its management and context of use; the "why" defines the purposes of analyzing the data collected (including monitoring, intervention, or reflection); the "how" defines the method to apply in the analysis of such data; and finally, the "who", the stakeholders to whom the analysis is directed [17].

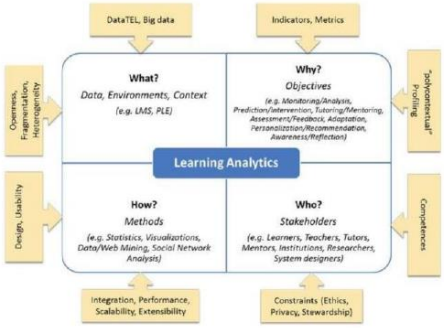

Fig. 1. Learning Analytics model dimensions [17].

Freire et al. [10] presented an abstract overview of a Game Learning Analytics (GLA) system, detailing all the steps of such architecture. Their described process starts when the game sends data to a collector. The data collected is aggregated to generate information to feed reports and visualizations (in real-time or offline) and assess students. The process ends in the adapter component, that provides feedback to adapt the game to players (Fig. 2).

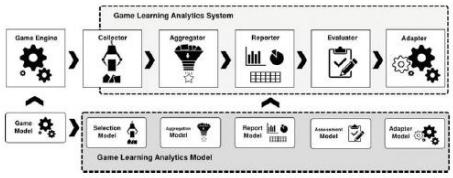

Fig. 2. Conceptual architecture of a Game Learning Analytics System, from [10].

The final step to turn the collected interaction data into usable and understandable information is to communicate it to the suitable stakeholders in a clear way. Visualizing the aggregated data, or the results of the performed analysis, in dashboards or visualizations could achieve this purpose. Lower-level analytics details can be hidden when required,

providing an overview of the results simplified for the relevant stakeholders (students, teachers, researchers). Many exploratory studies have created their dashboards, but more research is needed to compare dashboards designs. Learning Analytics dashboards display learners' data to make informed decisions about the learning process [18]. Examples of Learning Analytics dashboards have been developed, for instance, to support communication between students and advisers, to visualize learners' actions and GLA information [19]. Although the main goal of LA dashboards is to make learners aware of their learning process, authors have pointed out that, beyond that awareness, dashboards should also have aim to improve competencies.

The information obtained analyzing the collected interaction data can be integrated into the different phases of the videogame lifecycle and targeted to different user profiles with different goals. During development, the integration of such analyses is used to validate the design and find possible software errors. When validating the game, analytics helps to study the game's effect on players, as well as their interactions with the game. Finally, during deployment and application in real scenarios, analyses help teachers to assess players and track their progress during their sessions. Throughout this lifecycle of serious games, the nature and content of collected data can vary; large volumes of data will require analytics systems ready to process them, and the time and cost to create, configure and use those systems can easily be out of reach for many potential users. But before requiring a full analytics platform as may be required for truly large-scale deployments, during the SG development phase, the design and development team can benefit from a more agile approach – which we could term a "minimum viable analytics" (MVA) solution.

Increasing use of analytics is not only a problem of having a platform that can display it. Serious games must be created with analytics in mind, and the games must have an in-built mechanism to send analytics data for analysis. To reduce some of the barriers that currently exist in GLA, some platforms to simplify the collection and application of user interactions data have been developed: the serious games authoring tool uAdventure [20] integrates default learning analytics in its created games to simplify the definition and collection of such interaction data, while the validation tool SIMVA [21] aims to simplify the performance of experiments where interaction data is collected from serious games, providing an easy-to-use interface to then gather all the collected data.

In order to simplify the integration of analytics and to generalize their use and compatibility with other data sources one of the first steps is to use standards in the interaction data collected during the game sessions. These data standards should provide a clearly defined format to collect the interaction data, helping other researchers and users to clearly understand the information collected and simplifying integration with other tools and ecosystems. Besides the use of a standardized data format, privacy and anonymity requirements should be met, to comply with all applicable regulations (e.g. GDPR). One of the most widely used information standards in the educational field is xAPI. This standard allows the creation of specific profiles to adapt to the needs of the different educational resources that exist, such as serious games.

A. Experience API for Serious Games

The Experience Application Programming Interface (xAPI, for short) is a data specification created by a community led by the Advanced Distributed Learning (ADL) initiative, a program under the Department of Defense of the United States of America [22]. xAPI is based on activity streams, a standard to represent activities, and aims to provide a standard to communicate information about learners' activities in learning systems. The main concepts of xAPI are verbs, activity types and extensions. Data traces in xAPI (called *statements*) are JSON-based and represent learning activities. Each statement contains three main fields: *actor*, *verb*, and *object*. The *actor* represents the one who carries out the action, the *verb* is the action itself, and the *object* is the item that receives the action. Extensions may be included in the statements to provide further information about the learning activity such as: context, results, timestamp, etc.

For situations that have specific requirements that go beyond the ones defined in Experience API, specific xAPI Profiles can be created to provide the means to comply with expertise in that topic area. An xAPI Profile is defined as "the human or machine-readable documentation of application-specific concepts, extensions, and statement templates used when implementing xAPI in a particular context". xAPI Profiles provide a specific set of verbs, activity types and extensions to meet the needs of a specific area. Students' results in xAPI format can be stored in Learning Record Stores (LRSs). The data representation format is used to store the data in LRSs and to help transfer and combine data from multiple LRSs.

The xAPI Profile for Serious Games (xAPI-SG) was created to identify and standardize the common interactions that can be tracked in a serious game. An interaction model for serious games was created and then validated and published with ADL to be the official xAPI Profile for Serious Games [23]. The Profile defines a set of verbs (*accessed, completed, initialized, interacted, pressed, progressed, released, selected, skipped, unlocked, used*) and activity types (*area, controller, cutscene, dialog-tree, enemy, item, keyboard, level, menu, mouse, non-player-character, quest, question, screen, serious-game, touchscreen, zone*) that can be used to define the data from players' interactions in the game. This set of verbs and activity types covers the most common interactions that occur in serious games, including information about *completables* (game parts that can be started, progressed in and completed), or *accessibles* (game areas that can be entered and skipped). For instance, Fig. 3 depicts an example


```
{
  "actor": {
    "name": "John Doe",
    "mbox": "mailto:johndoe@example.com"
  },
  "verb": {
    "id": "https://w3id.org/xapi/adb/verbs/selected",
    "display": { "en-US": "selected" }
  },
  "object": {
    "id": "http://rage.e-ucm.com/activities/Countrix/questions/Capital_of_Spain",
    "definition": {
      "type:" "http://adlnet.gov/expapi/activities/question"
    }
  },
  "result": {
    "response": "Lisbon",
    "success": false,
    "extensions": {
      "https://w3id.org/xapi/seriousgames/extensions/health": 0.34
    }
  }
}
```


Fig. 3. Sample xAPI-SG statement capturing that the actor (John Doe) has selected a false response (Lisbon) in a question (Capital_of_Spain), and his current health is 0.34.

xAPI-SG statement representing that a player (given in the *actor* field), has selected (*verb* field) an incorrect response (given in the *response* and *success* fields of the *result*) in a question (*object* field). Using the xAPI-SG standard, we have developed T-Mon, a platform to simplify the analysis and visualization of interaction data from serious games.

III. T-MON: A PLATFORM TO SIMPLIFY AND AUTOMATE THE DATA ANALYSIS IN SERIOUS GAMES

T-Mon provides a default, game-independent set of analysis and visualizations to obtain information of serious games interaction data that follows the xAPI-SG standard. T-Mon contains a set of Jupyter Notebooks that process the xAPI-SG statements, analyzes them, and displays a default set of visualizations that provide a quick overview of its contents. All this process occurs automatically after the interaction data is loaded in T-Mon, providing an overview of the information collected in the data. The displayed information is useful to analyze the collected data and visualize the results of players' actions in the games.

The main Jupyter notebook in T-Mon expects a JSON file with the list of xAPI-SG traces to be processed. Certain xAPI traces, not specific to the xAPI-SG Profile, could also be processed by T-Mon; but the analysis mainly focuses on the specifics of the Serious Games Profile. The traces in the JSON file are then analyzed by T-Mon. The xAPI-SG traces are read in order and processed individually. For each player (given in the *actor* field of the traces), T-Mon stores a set of higher-level game learning analytics information, creating a set of variables that constitute the player profile. The information on each player profile is updated with each subsequent trace corresponding to the same player. The information for each player is stored in higher-level metrics that differentiate the information gathered for each type of verb included in the Profile: *initialized, completed, progressed, accessed, skipped, interacted* and *selected*.

The default set of visualizations is then filled with the information aggregated in each player profile. T-Mon's interface displays the results in visualizations grouped in 7 tabs containing information about: players' progress, use of videos, completables, alternatives, interactions with items, accessibles and menus (Fig. 4). We currently provide default game-independent visualizations with the following information:

- Start, completion and progress of players in the SG
- Final progress in completables, with evolution over time
- Final scores obtained in completables
- Maximum and minimum completion time in completables
- Correct and incorrect responses in alternatives per player, and per alternative
- Responses selected in questions (alternatives)
- Interactions and actions with items
- Videos (accessibles) seen and skipped
- Accessibles accessed
- Selections in menus

Fig. 5 and Fig. 6 display some of T-Mon's default visualizations, populated with sample xAPI-SG. The plot style

Fig. 4. T-Mon configuration options. From top to bottom: button to upload the xAPI-SG file, plot style dropdown, the seven visualization tabs, selection of players and other data (e.g. completables).

of the visualizations can be changed using a drop-down menu. The data displayed can be modified by selecting or removing specific player data from the visualizations, or specific items. For example, in the *completables* tab, specific completables can be selected, filtering the corresponding view so that data from non-selected completables is filtered out. These configuration options are displayed in Fig. 4. Visualizations can be further configured by: selecting whether data should be displayed in absolute values or as percentages, selecting the number of items to appear per visualization (if there is too much data, this can be divided into multiple visualizations), and ordering the data in the x axis (in alphabetical order, from higher to lower values or from lower to higher values). Visualizations with information per time can be configured to be displayed in absolute time or relative time (that is, relative to the first data point for each player, to compare between sessions carried out in different dates).

To expand the functionality of T-Mon, we have configured integration with SIMVA (which stands for Simple Validator), a tool to simplify experiments to validate and deploy serious games [21]. SIMVA manages the commonly used questionnaires, as well as the interaction data, storing all results, and linking all data from each player using anonymous identifiers. Integration between T-Mon and SIMVA allows interaction data collected from experiments with serious games in SIMVA to be accessed seamlessly from T-Mon. The default analysis and visualizations available are then applied to the data as provided by SIMVA in xAPI-SG format.

T-Mon uses some common Python libraries to perform the analysis and visualizations. Apart from these, T-Mon does not require any further configuration, as all analysis and visualizations are performed and displayed automatically. This way, the tool is accessible to non-experts in the domain. Additionally, data scientists can perform further analysis in the Python Jupyter Notebooks to extend the analysis and visualizations included. T-Mon is openly and freely available on GitHub[1], to be downloaded and launched locally (Fig. 7). Additionally, T-Mon can also be launched remotely using Binder (directly from the GitHub repository). The Binder launching deals with all library dependencies and provides a web-based interface to test the tool uploading the xAPI-SG data file.

[1] https://github.com/e-ucm/t-mon

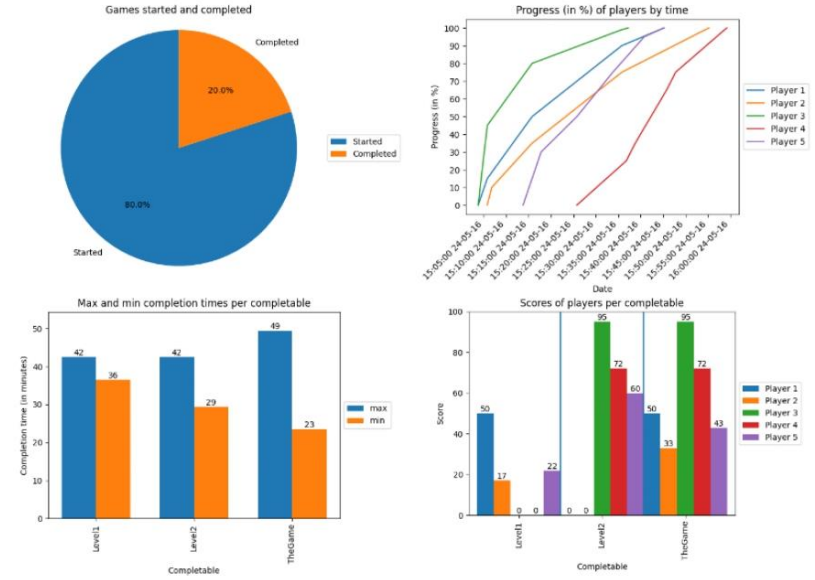

Fig. 5. Four of the default visualizations included in T-Mon (left to right, top to bottom): pie chart with percentage of serious games started and completed; line chart with progress (*y-axis*) of each player in the game over time (*x-axis*); bar chart with maximum and minimum completion times (*y-axis*) in each completable (*x-axis*), max and min times corresponding to each bar per completable; and bar chart with scores (*y-axis*) obtained by each player in each completable (*x-axis*), each bar per completable corresponding to one player.

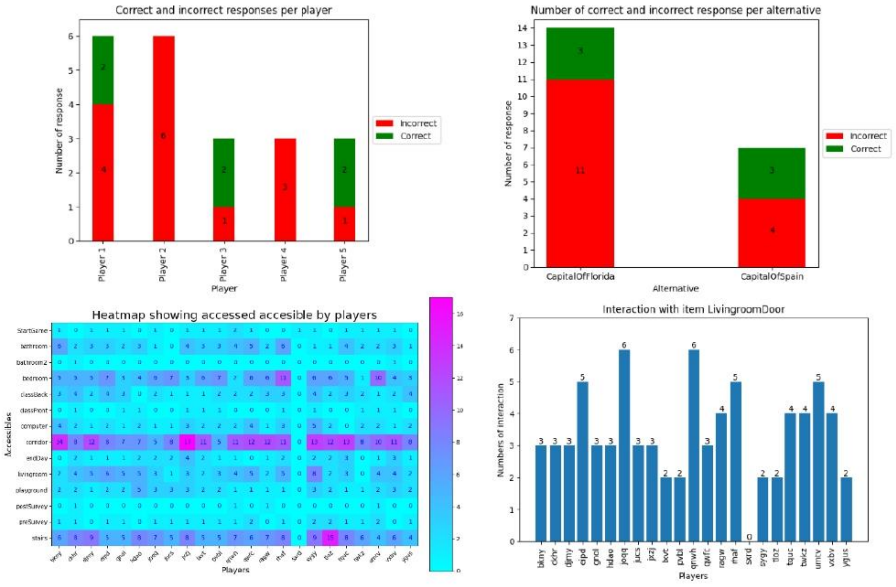

Fig. 6. Four of the default visualizations included in T-Mon (left to right, top to bottom): bar chart with correct (in green) and incorrect (in red) number of responses (*y-axis*) in alternatives per player (*x-axis*); bar chart with correct (in green) and incorrect (in red) number of responses (*y-axis*) per alternative (*x-axis*); heatmap with times each accessible (*y-axis*) has been accessed per player (*x-axis*); and bar chart with number of interactions (*y-axis*) per player (*x-axis*) with an item.

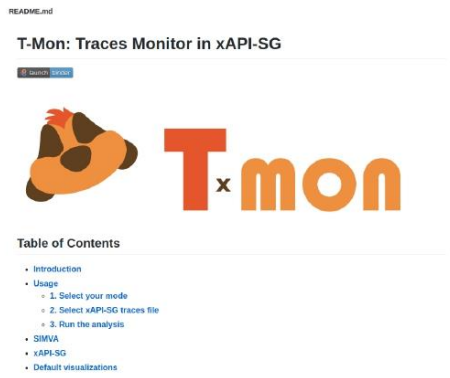

Fig. 7. T-Mon GitHub page, available at https://github.com/e-ucm/t-mon.

IV. Discussion and Conclusions

To extract information from players' interactions in serious games, the whole process needs to be considered: from the definition of the relevant interactions in the game, their collection in a specific format, and the aggregation, analysis and reporting of the information gathered. For this process to provide meaningful results, the relevant data to be collected from the games needs to be clearly defined from the very start [24]. The use of a standardized data format to collect the user interaction data and the use of a pre-defined language to represent the interactions and specific game mechanics allow the creation of tools to automate analysis, such as T-Mon. The development of tools that simplify data collection and analysis, as well as the use of standards and clearly-defined data models to integrate analytics in serious games, can ease the study, evaluation and adoption of games as educational tools, minimizing the current barriers for their adoption. These tools simplify the application of serious games for users with a less technical profile: T-Mon helps to build the minimum viable analytics, by providing a low-cost entry point to the use of game learning analytics and the validation of serious games. This could also be the case of data-science experts that have little game analytics knowledge. The use of standards is also clear benefit to simplify later analysis, and integration with other ecosystem services and tools [25]. Additionally, if allowed by the data management guidelines, the use of the standard also simplifies data sharing, providing a clear framework to understand the interaction data shared.

Once data is collected, to actually provide meaningful information to the stakeholders involved, it is useful to provide some visual display of aggregated data or results, hiding the low-level details about the analysis performed and, instead, showing meaningful summaries of gathered information. The application of tools that provide default game-independent analysis and visualizations detaches users from the details of the analysis required. This simplifies their use, as they can be adopted as a black box, and the final user does not need to know the characteristics of the collection data format or the data analysis. This is the case of the default set of analysis and visualizations provided by T-Mon. Even more so, by using the xAPI-SG standard, the reports provided can be obtained without any knowledge of the game design details, isolating the analysis and visualizations from the game design. This could benefit, among others, game experts with little knowledge of data analysis. With this features, T-Mon covers both the steps of statistics and visualizations considered in Learning Analytics models [17] as well as the reporting and visualizations (including offline) of results considered in Game Learning Analytics models [10]. All this information can be obtained remotely using the tool available online, which simplifies its application in diverse contexts, such as the ones faced currently in the New Normal after the covid-19 pandemic.

The level of detail and granularity in the result visualizations is adaptable according to the amount and characteristics of the interaction data collected. T-Mon allows simple descriptive analysis of player decisions, but also allows performing more detailed and complex default analysis. For instance, a possible lower-level analysis would entail examining in-game conversations, to determine whether players are actually taking the time to read conversation lines or simply skipping them. By collecting interaction data of the relevant actions in conversations (e.g. as accessibles or completables), T-Mon could display the information of such players' actions to see if players are skipping conversations, or how much time they are spending in them.

Additionally, T-Mon can also be used by data scientists and other technical users, who could extend and complement the default set of analysis and visualizations if needed to meet any requirements that go beyond the ones covered by default. This includes the possibility of exploring more complex data mining and machine learning techniques (e.g. for predictions) which, given knowledge about the game and learning design, could complement the default analysis to provide a more evidence-based assessment of players based on their game decisions, further extending their scope of application. For non-experts, however, the ready-to-use analysis and visualizations provide an overview of the interaction data to extract information about players' progress and process in the serious game.

T-Mon provides a default set of analysis and visualizations that can be used to report and visualize results of the interaction data collected from serious games, using the xAPI-SG standard. Data that adheres to the standard is analyzed, and all the fields and types included in the SG Profile are used in the default analysis and visualizations. T-Mon can therefore help to easily and quickly obtain an overview of the interaction data collected from the serious game. T-Mon's simple and user-friendly interface and the data standard used can further simplify integration with other systems. With this tool, and the only requirement of using the xAPI-SG standard data format, we expect to simplify the analysis and visualization of interaction data from serious games.

V. Limitations and Future Work

The tool has some limitations: the requirement that the input data should follow the xAPI-SG standard limits its application to other types of interaction data collected. However, this standard data format is broad and flexible enough so that most serious game interactions could be tracked using this format and, therefore, be analyzed and visualized using T-Mon. The analysis and visualizations included could be further extended with two perspectives: on the one hand, providing more in-depth information about some of the specific types included in the Profile; on the other, including analysis and visualizations that are more general to xAPI information that do not meet the specifics of the xAPI-SG Profile: we plan to continue working on the tool to further extend the analysis and visualizations included.

For the moment T-Mon has so far been tested and improved with xAPI-SG data collected in several previous experiments of the research group. In the future, we will carry out new case studies with other researchers to further evaluate its usability and improve the tool.

Acknowledgment

This work has been partially funded by Regional Government of Madrid (eMadrid S2018/TCS-4307, co-funded by the European Structural Funds FSE and FEDER), by the Ministry of Education (TIN2017-89238-R), and by the European Commission (Erasmus+ IMPRESS 2017-1-NL01-KA203-035259).

The authors would like to thank Julio Santilario Berthilier for his contributions to the tool.

References

[1] D. R. Michael and S. L. Chen, "Serious Games: Games That Educate, Train, and Inform," *Education*, vol. October 31, pp. 1–95, 2005.

[2] D. Djaouti, J. Alvarez, and J.-P. Jessel, "Classifying Serious Games," in *Handbook of Research on Improving Learning and Motivation through Educational Games*, no. 2005, IGI Global, 2011, pp. 118–136.

[3] J. P. Gee, "What video games have to teach us about learning and literacy," *Comput. Entertain.*, 2003.

[4] D. B. Clark, E. Tanner-Smith, A. Hostetler, A. Fradkin, and V. Polikov, "Substantial Integration of Typical Educational Games Into Extended Curricula," *J. Learn. Sci.*, vol. 8406, no. June, p. 10508406.2017.1333431, 2017.

[5] L. Jean Justice and A. D. Ritzhaupt, "Identifying the Barriers to Games and Simulations in Education: Creating a Valid and Reliable Survey," *J. Educ. Technol. Syst.*, vol. 44, no. 1, pp. 86–125, Sep. 2015.

[6] A. Calderón and M. Ruiz, "A systematic literature review on serious games evaluation: An application to software project management," *Comput. Educ.*, vol. 87, pp. 396–422, Sep. 2015.

[7] S. de Klerk and P. Kato, "The Future Value of Serious Games for Assessment: Where Do We Go Now?.," *J. Appl. Test. Technol.*, vol. 18, no. February, pp. 32–37, 2017.

[8] P. Long and G. Siemens, "Penetrating the Fog: Analytics in Learning and Education," *Educ. Rev.*, vol. 46, no. 5, pp. 30–32, 2011.

[9] M. Bienkowski, M. Feng, and B. Means, "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," *Washington, DC SRI Int.*, pp. 1–57, 2012.

[10] M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, "Game Learning Analytics: Learning Analytics for Serious Games," in *Learning, Design, and Technology*, Cham: Springer International Publishing, 2016, pp. 1–29.

[11] C. S. Loh, Y. Sheng, and D. Ifenthaler, *Serious Games Analytics*. Cham: Springer International Publishing, 2015.

[12] V. Shute and M. Ventura, "Stealth Assessment," in *The SAGE Encyclopedia of Educational Technology*, 2455 Teller Road, Thousand Oaks, California 91320: SAGE Publications, Inc., 2013, p. 91.

[13] Z. Papamitsiou and A. A. Economides, "Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence," *Educ. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, 2014.

[14] O. Adejo and T. Connolly, "Learning Analytics in Higher Education Development: A Roadmap," *J. Educ. Pract.*, 2017.

[15] Y. Chaudy, T. Connolly, and T. Hainey, "Learning Analytics in Serious Games: a Review of the Literature," *Ecaet 2014*, no. March 2016, 2014.

[16] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Applications of data science to game learning analytics data: A systematic literature review," *Comput. Educ.*, vol. 141, p. 103612, Nov. 2019.

[17] M. A. Chatti *et al.*, "Learning Analytics: Challenges and Future Research Directions," *E-Learning Educ.*, no. 10, 2015.

[18] I. Jivet, M. Scheffel, M. Specht, H. Drachsler, and M. Specht, "License to evaluate: Preparing learning analytics dashboards for the educational practice," in *Learning Analytics & Knowledge Conference*, 2018, pp. 31–40.

[19] A. Calvo-Morata, C. Alonso-Fernández, I. J. Perez-Colado, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Improving Teacher Game Learning Analytics Dashboards through ad-hoc Development," *J. Univers. Comput. Sci.*, vol. 25, no. 12, pp. 1507–1530, 2019.

[20] V. M. Pérez Colado, I. J. Pérez Colado, F. Manuel, I. Martínez-Ortiz, and B. Fernández-Manjón, "Simplifying the creation of adventure serious games with educational-oriented features," *Educ. Technol. Soc.*, vol. 22, no. 3, pp. 32–46, 2019.

[21] I. J. Perez-Colado, A. Calvo-Morata, C. Alonso-Fernandez, M. Freire, I. Martinez-Ortiz, and B. Fernandez-Manjon, "Simva: Simplifying the Scientific Validation of Serious Games," in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, 2019, pp. 113–115.

[22] ADL, "Experience API," 2012. [Online]. Available: https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/. [Accessed: 20-Mar-2016].

[23] Á. Serrano-Laguna, I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, and B. Fernández-Manjón, "Applying standards to systematize learning analytics in serious games," *Comput. Stand. Interfaces*, vol. 50, pp. 116–123, 2017.

[24] K. Kitto, J. Whitmer, A. E. Silvers, and M. Webb, "Creating Data for Learning Analytics Ecosystems," *Sol. Position Pap.*, pp. 1–43, 2020.

[25] M. Liu, J. Kang, S. Liu, W. Zou, and J. Hodson, "Learning Analytics as an Assessment Tool in Serious Games: A Review of Literature," in *Serious Games and Edutainment Applications*, Cham: Springer International Publishing, 2017, pp. 537–563.

6.2.3. Full lifecycle architecture for serious games: integrating game learning analytics and a game authoring tool

Full citation

Cristina Alonso-Fernández, Dan C. Rotaru, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2017): **Full Lifecycle Architecture for Serious Games: Integrating Game Learning Analytics and a Game Authoring Tool.** Joint Conference on Serious Games (JCSG), 23-24 November 2017, Polytechnic University of Valencia, Spain.

Abstract

The engaging and goal-oriented nature of serious games has been proven to increase student motivation. Games also allow learning assessment in a non-intrusive fashion. To increase adoption of serious games, their full lifecycle, including design, development, validation, deployment and iterative refinement must be made as simple and transparent as possible. Currently serious games impact analysis and validation is done on a case-by-case basis. In this paper, we describe a generic architecture that integrates a game authoring tool, uAdventure, with a standards-based Game Learning Analytics framework, providing a holistic approach to bring together development, validation, and analytics, that allows a systematic analysis and validation of serious games impact. This architecture allows game developers, teachers and students access to different analyses with minimal setup; and improves game development and evaluation by supporting an evidence-based approach to assess both games and learning. This system is currently being extended and used in two EU H2020 serious games projects.

Full Lifecycle Architecture for Serious Games: Integrating Game Learning Analytics and a Game Authoring Tool

Cristina Alonso-Fernandez[1], Dan C. Rotaru[1], Manuel Freire[1], Ivan Martinez-Ortiz[1] and Baltasar Fernandez-Manjon[1]

[1] Facultad de Informática, Complutense University of Madrid, C/ Profesor José García Santesmases 9, 28040 Madrid, Spain
{crisal03, drotaru}@ucm.es
{manuel.freire, imartinez, balta}@fdi.ucm.es

Abstract. The engaging and goal-oriented nature of serious games has been proven to increase student motivation. Games also allow learning assessment in a non-intrusive fashion. To increase adoption of serious games, their full lifecycle, including design, development, validation, deployment and iterative refinement must be made as simple and transparent as possible. Currently serious games impact analysis and validation is done on a case-by-case basis. In this paper, we describe a generic architecture that integrates a game authoring tool, uAdventure, with a standards-based Game Learning Analytics framework, providing a holistic approach to bring together development, validation, and analytics, that allows a systematic analysis and validation of serious games impact. This architecture allows game developers, teachers and students access to different analyses with minimal setup; and improves game development and evaluation by supporting an evidence-based approach to assess both games and learning. This system is currently being extended and used in two EU H2020 serious games projects.

Keywords: Learning Analytics, Serious Games, E-learning, xAPI, uAdventure

1 Introduction

Games have been applied in multiple fields such as medicine [1], science [2], arts [3] or military [4]. Their benefits, such as their goal-oriented, engaging nature, makes them especially adequate for education, where students' motivation is essential.

Serious Games (SGs) are videogames where the main purpose is not pure entertainment: it may be to teach, to change an attitude or behavior or to create awareness of a certain issue [5]. There are several examples of successfully applied SGs: *Aislados* helped teenagers to prevent drug addiction and other risk behaviors [6] while *Darfur is Dying* created awareness of the ongoing war in Sudan in 2006 [7].

Most games, however, follow the black box model when it comes to collecting players' interactions: they merely report final results, which are far less informative

than access to real-time learning progress. In fact, the usual method to evaluate SGs effectiveness is through pre-post questionnaires [8]. This evaluation method requires significant investments of time and effort, and individual solutions have to be provided ad hoc for every particular game, severely impacting the scalability of the solution.

The pre-post evaluation method also fails to detect changes in learning as they occur. Learning concepts appear at different stages of the game for different players; and this learning process should be tracked in real-time through the observation of in-game interactions for optimal feedback regarding the effectiveness of the games' learning design.

In the entertainment games industry, data analysis has been long applied to capture players' interactions and to improve their user experience as well as the game design [9] in a discipline that is usually called Game Analytics (GA). Meanwhile,

In e-learning and different learning systems, such as learning management systems (LMS), Learning Analytics (LA) is commonly used to capture learners' actions to try to understand their learning process and prevent their failure. There are several definitions of LA; we could define it as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which it occurs" [10]-[11].

To apply these analytics models to SGs, we have to break the black box model to gain insights of players' interactions as they take place. This information could be then related to each student's learning process. Game developers could also benefit from this information to determine areas of minor or greater difficulty for players, or even game bugs such as unreachable areas. If the game design is suitable and the relevant interaction data is captured by tracing players' interactions in the SG, it should be possible to trace the evolution of their knowledge, telling apart the areas where they struggle or shine.

In Section 2, we describe the lifecycle of evidence-based SGs' impact. In Section 3, we describe the proposed abstract architecture for applying game and learning analytics for SGs and the different steps it comprises in design, development and evaluation. In Section 4, we describe a reference implementation as part of two EU H2020 SG-related projects. Finally, in Section 5 we summarize the main contributions and future work.

2 Lifecycle of evidence-based serious games' impact

The combination of LA methods with the technologies long applied in GA allows players' interactions within SGs to be traced and analyzed, providing insight into their learning progress. We call this process Game Learning Analytics [5]. GLA allows an evidence-based approach to games' lifecycle (e.g. development, validation and evaluation).

The lifecycle of a serious game (see Fig. **1**) goes from initial conception to development, validation (which may require several iterations if design flaws are uncovered before widespread release), and exploitation, during which periodic

evaluation of student progress and outcomes will take place, once the game is released to its target players. The role of integrated analytics is critical to collecting and analyzing interactions to generate actionable feedback.

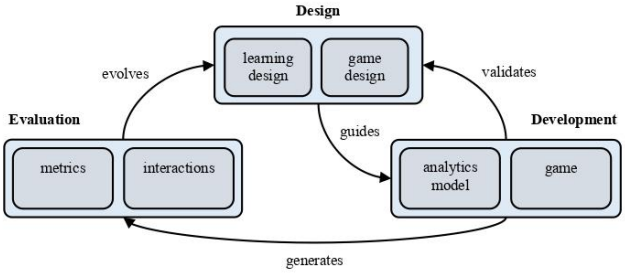

Fig. 1. Lifecycle of a serious game: from learning and game design, through development, validation and evaluation.

Both validation and evaluation require a strong integration of analytics to benefit from feedback and allow players to be evaluated meaningfully on their progress. The use of a unified analytics framework that can do GLA (that is, analyses both at the game level and the learning level) combining separate systems presents advantages for both. Integration of GLA into the development platform also presents significant benefits, comparable to those that test-driven design brings to programming: an early emphasis on choosing and measuring evidence of quality.

Both teachers and students can benefit from closer analytics integration. Analytics can provide real-time knowledge of what students are doing, but interpreting the data is difficult unless it is well presented. Dashboards that combine complementary visualizations appear to be an appropriate way of communicating data to stakeholders, who generally do not need to understand the details of the analysis performed underneath.

To achieve the most informative results, ad-hoc visualizations would be needed; however, providing meaningful default dashboards ensures that no setup is required to start enjoying advantages.

3 Proposed abstract architecture

We propose a complete and scalable analytics architecture, based on standards, that encompasses the whole process from game development to the analysis and visualization of results; a design guided game development where the interaction tracker sends players' interactions to the analytics platform composed of collector,

analysis and dashboard; feedback will be sent back to the learning and game design (see Fig. **2**).

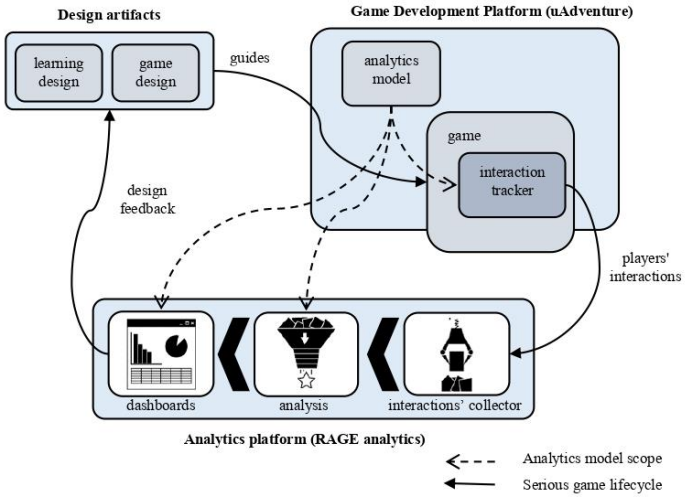

Fig. 2. Proposed architecture: design guides game development.

In this architecture, SGs send traces in a standard-based format to a server that analyzes the data and transforms it into useful information. This analyzed information is then displayed in dashboards for different stakeholders: teachers or instructors in charge of players, the players themselves, game developers or designers and researchers [12].

3.1 Analytics models

Essential questions when performing analytics are those of what to track, how to analyze it, and how to present the results. Another way to frame these questions is to attempt to create a list of visualizations that would evidence that the goals of the game mechanics and the learning design are being met, and work backwards to define the analyses and data-collection that would be required. As illustrated in Fig. **2**, analytics models inform all these decisions, and are an integral part of the game development.

The best analytics models are those that are designed together with the game itself, and are both influenced by the game's design and, where necessary, result in changes to the design that make the resulting game easier to analyze. However, there is a

strong case for providing default analytics models whenever possible, to minimize the burden on game designers and developers (which could otherwise decide to forgo GLA altogether) and provide sane defaults on which more targeted analytics can be built.

3.2 Interaction tracking

Analytics requires collection of each player's interaction with the game prior to any analysis. A standard collection format is desirable to allow interoperability and avoid data lock-in. After analyzing the current state of data standards and SGs, in addition to previous experiences applying e-learning standards to SGs [13]-[14], a new interaction representation model has been defined and implemented based on the Experience API standard [15]-[16]: the xAPI Serious Games vocabulary, or xAPI-SG for short.

Experience API (xAPI) is a data format developed by a community led by the Advanced Distributed Learning Initiative (ADL) [17]. The standard derives from Activity Streams, which represent a series of statements regarding learning activities with three main attributes: an actor, a verb and an object. Additional attributes may be included such as the result of the action or a timestamp. Fig. **3** shows an xAPI-SG sample trace generated with [18] representing that the learner completed the SG with a score of seven.


```
{
    "actor": {
        "mbox": "mailto:learner@example.com",
        "name": "Example Learner",
        "objectType": "Agent"
    },
    "verb": {
        "id": "http://adlnet.gov/expapi/verbs/completed",
        "display": {
            "en-US": "completed"
        }
    },
    "object": {
        "id": "http://adlnet.gov/expapi/activities/serious-game",
        "definition": {
            "name": {
                "en-US": "Serious Game"
            },
            "description": {
                "en-US": "Serious game example"
            }
        },
        "objectType": "Activity"
    },
    "result": {
        "score": {
            "raw": 7
        }
    }
}
```


Fig. 3. Sample xAPI trace. The serious game activity was completed by Example Learner with result score of 7.

The interaction model comprises several concepts such as *completables* (e.g. levels, quests or the serious game), *alternatives* (e.g. options in questions or dialogs) and general *variables* to track interactions in the specific domain of SGs. If desired, custom interactions may also be defined to extend the information for a specific SG.

In the architecture illustrated in Fig. **2**, games contain an interaction tracker component that communicates players' in-game interactions via xAPI-SG to the analytics platform. The analytics model defines which interactions, events and targets are reported and how they are mapped to their corresponding xAPI-SG statement attributes, verbs and activity types.

3.3 Game Development Platform

Integrating tracking of GLA with a developed SG is typically performed ad hoc, and both the tracker and the analytics model are external to the chosen game development platform. However, a game development platform which follows the architecture illustrated in Fig. **2**, must include the tracking component in each game, and configure it with an analytics model that is fully integrated with the game's authoring environment. This integration greatly reduces the investment of time and effort required from game developers to benefit from analytics, and therefore increases the likelihood that they will be able, with some additional effort, to improve the game design, the analytics model, and most importantly the game itself in each successive iteration of its lifecycle.

In our reference implementation of the architecture this component is implemented using uAdventure [19]-[20], a complete rewrite of eAdventure, an authoring tool for point-and-click games written in Java and previously developed by the e-UCM Research Group [21]-[22]. As many platforms and devices no longer support Java, uAdventure is built on Unity3D.

3.4 Data analysis

Once the data is collected, the analytics server can begin to process it. Again, the analytics model must provide information on the metrics and KPIs that will be used to prove the effectiveness of the learning design. We distinguish two types of analysis:

1. Game-independent analysis that should be suitable for any SG that connects to the analytics server, as long as the game generates standards-compliant xAPI-SG traces.
2. Game-dependent analysis, which must be developed ad hoc for each game, but allow game and learning designers to create dashboards that perfectly match their game's goal and design.

The information obtained as result of the analysis should be stored for its later visualization; in the proposed architecture, analytics results are stored in a time series database (in our reference implementation of the architecture, Elasticsearch [23] is

.

used), which can analyze and query large amounts of data in semi-real time, and is especially suited for later visualization.

3.5 Data visualization

To facilitate the understanding of the analysis results for a range of stakeholders (including teachers, students and developers), it is important to provide each with informative dashboards to display results. The need for easy to understand and informative visualizations is especially important in the case of teachers, which can greatly benefit from real-time information to monitor a class while students are playing a game, and to provide targeted feedback to students that get stuck. Students will see their own personal progress in real-time and a general ranking within the same class competing with other students. Visualizations about the overall usage of the games such as session's length and server loading are shown to the developers, though the student specific data is only shown to the teacher because of privacy concerns.

In our reference implementation of the architecture of this component, visualization dashboards have been developed using Kibana [24], an open source visualization engine which is directly connected with Elasticsearch. Kibana provides a browser-based interface to quickly develop analysis and visualizations with different predefined graphics (e.g. line chart, bar chart, pie chart). Two sample visualizations available: the left one shows number of correct (in green) and incorrect (in red) answers in each alternative; the right one, the progress (in range 0 to 1) in each of the three completables and in the complete serious game for each player (see Fig. **4**). New visualizations and dashboards may be configured in the system by selecting the required fields to be analyzed and displayed in the graphs.

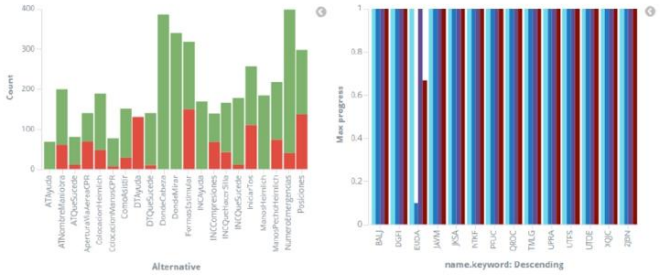

Fig. 4. Two sample visualizations containing errors made in *alternatives* and progress made in *completables*.

To extend the usefulness of these visualizations, recommended actions may be included to help teachers provide timely feedback. In our implementation, alerts (situations that require immediate action, e.g. "a student has made an important

error") and warnings (less urgent actions, e.g. "a student has been inactive for two minutes") provide near real-time information to teachers. Fig. **5** shows the general view of alerts and warnings (on top); clicking on a specific user in the general view displays details on the selected user's alerts and warnings (bottom part).

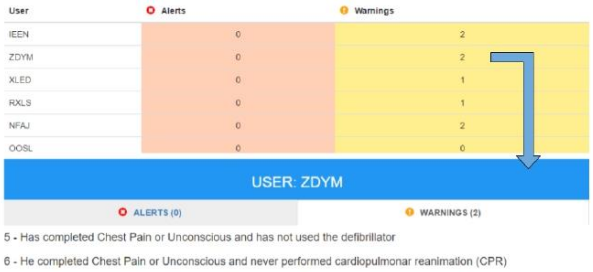

Fig. 5. General view of alerts and warnings for each anonymized user. Clicking on a specific user provides further details on the user's alerts and warnings.

4 GLA architecture reference implementation

A complete architecture to manage GLA requires handling several interlocking parts: data tracking, data analysis and results visualization. We proposed the following standard-based architecture, a combination of modules that work together to analyze and visualize information collected from SGs [12]. Fig. **6** shows a diagram of the GLA architecture: from learning and game design, the serious game is created. Its embedded tracker sends xAPI traces to the collector, which stores them in a LRS for batch analysis and sends them for real-time analysis. Visualizations developed from analytics provide feedback to come full circle improving the learning and game design and helping to assess students.

- The learning and game design determine the SG implementation. This includes the game mechanics, structure, goals and in-game items or characters. Both these designs also determine the elements that will contain the relevant information for learning (usually as game variables), that is, the elements that are essential to be tracked as they will tell if the game is helping players to learn or not.
- The SG itself will use a tracker component to send traces in xAPI format (called *statements*). The tracker provides an application programming interface designed to send data to the server without having to know the underlying xAPI format specification. Current tracker implementations include Unity C#, pure C# and JavaScript to facilitate their integration with different SGs.
- xAPI-SG statements are sent to a collector endpoint on the server-side. Then they are sent to a real-time analysis component which updates the information for each

player. xAPI statements are also sent to a Learning Record Store (LRS), a scalable database designed for storing xAPI learning activity data to perform a different analysis ("batch analysis").

- With the analysis results, visualizations in suitable dashboards show all metrics of interest for the stakeholders. If configured, personalized visualizations, alerts and warnings will also be displayed.
- Finally, the process is completed when the information obtained through analysis and visualizations provides feedback and improvement actions that can be reintroduced in the system for following iterations of the learning and game design. Additionally, this information may also help teachers in students' assessment.

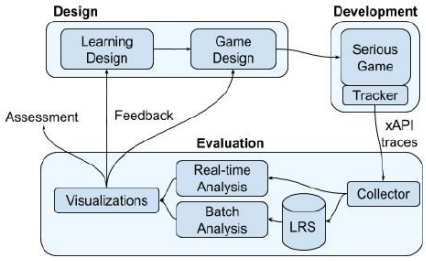

Fig. 6. GLA system: from design to development and evaluation.

4.1 Architecture implementation

The GLA architecture described above has been developed as part of an EU H2020 SG-related project. All components are open source and available online[1]. When deployed, the components are launched as Docker containers [25] which eases deployment by eliminating all dependencies except for Docker itself. The main components are:

- An authorization and authentication component (A2), which enforces access controls and allows integration with existing institutional single sign-on systems, as well as hiding the complexity of all other components behind a reverse proxy.
- A frontend that allows stakeholders (teacher, developer and student) to configure and/or view dashboards for which they have appropriate credentials.
- A backend that collects incoming traces, analyzes them (either in real-time from incoming data or on-demand from LRS queries), and exhibits results for the frontend.
- A xAPI Learning Record Store (LRS) which allows third-party systems to query xAPI traces collected by the backend.

[1] eUCM Research Group, RAGE Analytics, (2017). https://github.com/e-ucm/rage-analytics

Since teachers, students and other institutional stakeholders typically already enjoy single sign-on in institutional systems, the A2 component has been extended to interact with these institutional systems via login plugins for either SAML2 (Security Assertion Markup Language v 2.0) [26] or LTI (Learning Tools Interoperability) [27]. These plugins simplify deployment in applicable institutions, since no additional credentials need to be created. Additionally, in the case of LTI, certain setup tasks, such as registering students as belonging to a particular teacher's class, can be eliminated altogether. Further information about the RAGE Analytics System can be found in the e-UCM Research Group's GitHub wiki page [28].

4.2 System applications

The proposed system has been recently tested in an experiment with more than 200 students of a school in Madrid. With the goal of evaluating the whole GLA architecture, students played a SG to teach first aid maneuvers [29] while teachers obtained real-time dashboards about what students were doing, being able to control which students were progressing and which students were falling behind.

Students' anonymization was ensured via unique codes provided at the beginning of the experiment and required to access the game. Teachers were the only holders of the mapping between individuals and codes; all information collected in the system was only identified through this code.

During the experiments, it came to light that some visualizations were not easy to understand by teachers. As teachers are the only experts qualified in the evaluated field to know if students are learning or not, dashboards need to provide information in a clearer or more simplified manner for their easy comprehension.

5 Conclusions and future work

Although GLA is no longer an emerging field, it is still performed mostly through ad-hoc solutions, and therefore it could greatly benefit from a general standardized approach. Such an approach can increase adoption of SGs by promoting quality through evidence-based iterative improvement and better evaluation; while minimizing GLA deployment and development costs.

Our approach has three main pillars: first, the integration of analytics into the game authoring tool itself; second, the use of a standard xAPI-SG interaction model to standardize trace collection; and third, a default set of analysis and visualizations for the main SG stakeholders, including game developers, teachers and students.

Games created with the authoring tool uAdventure can effortlessly integrate tracking and analysis of results. Moreover, they can be deployed on a wide range of platforms, and can also support geolocalization [30].

Ad hoc analyses and visualizations can also be created by adding configuration files to the system or selecting the attributes to be visualized, respectively. Theses personalized analyses and visualizations could be useful if a particular game requires them; however, a moderate use of these is recommended as the more personalized

configurations included, the less general the solution will be and the more effort it will require.

With these contributions, we have advanced towards a systematized standards-based system that helps to complete the full circle of GLA for SGs: learning and game design, SG development, tracking, analysis, visualization, and feedback, as depicted in Fig. **2**.

However, there is still work to do. Some areas for improvement include:

- Improved explanations to allow novice users to interpret dashboard visualizations; especially for users that may not have been involved in the game design process.
- Simplified creation for custom visualizations. We are developing a wrapper around Kibana's built-in authoring environment to ease the process for non-programmers.
- Bidirectional communication between the tracker and the server, allowing the tracker to be notified when certain conditions are fulfilled in order to adapt the game's learning design and/or provide in-game, real-time feedback to players.

The system will be tested in more experiments with serious games currently under development. Work will continue on these and other improvements as the system is going to be improved and extended as part of the H2020 SG-related projects RAGE and BEACONING.

Acknowledgements

The e-UCM research group has been partially funded by Regional Government of Madrid (eMadrid S2013/ICE-2715), by the Ministry of Education (TIN2013-46149-C2-1-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676).

References

1. Evans, K.H., Daines, W., Tsui, J., Strehlow, M., Maggio, P., Shieh, L.: Septris: a novel, mobile, online, simulation game that improves sepsis recognition and management. Acad. Med. 90, 180–4 (2015).
2. Center for Game Science at the University of Washington: Treefrog Treasure, http://centerforgamescience.org/blog/portfolio/treefrog-treasure/.
3. Manero, B., Torrente, J., Serrano, Á., Martínez-Ortiz, I.: Can educational video games increase high school students' interest in theatre? Comput. 87, 182–191 (2015).
4. United States Army: America's Army, https://www.americasarmy.com/.
5. Freire, M., Serrano-Laguna, Á., Iglesias, B.M., Martínez-Ortiz, I., Moreno-Ger, P., Fernández-Manjón, B.: Game Learning Analytics: Learning Analytics for Serious Games. In: Learning, Design, and Technology. pp. 1–29. Springer International Publishing, Cham (2016).
6. Asociación Servicio Interdisciplinar de Atención a las Drogodependencias (SIAD): Aislados, http://www.aislados.es/zona-educadores/.
7. interFUEL, L.: Darfur is Dying, http://www.gamesforchange.org/play/darfur-is-dying/.

8. Calderón, A., Ruiz, M.: A systematic literature review on serious games evaluation: An application to software project management. Comput. Educ. 87, 396–422 (2015).
9. El-Nasr, M., Drachen, A., Canossa, A.: Game analytics maximizing the value of player data. Springer, London ;;New York : (2013).
10. Long, P., & Siemens, G.: Penetrating the Fog: Analytics in Learning and Education. Educ. Rev. 31–40 (2011).
11. Siemens, G., Dawson, S., & Lynch, G.: Improving the Quality and Productivity of the Higher Education Sector. Policy and Strategy for Systems-Level Deployment of Learning Analytics. (2013).
12. Alonso-Fernández, C., Calvo Morata, A., Freire, M., Martínez-Ortiz, I., Fernandez-Manjon, B.: Systematizing game learning analytics for serious games. In: IEEE Global Engineering Education Conference (EDUCON). pp. 1106–1113 (2017).
13. Serrano, A., Marchiori, E.J., del Blanco, A., Torrente, J., Fernandez-Manjon, B.: A framework to improve evaluation in educational games. In: Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON). pp. 1–8. IEEE (2012).
14. del Blanco, Á., Marchiori, E.J., Torrente, J., Martínez-Ortiz, I., Fernández-Manjón, B.: Using e-learning standards in educational video games. Comput. Stand. Interfaces. 36, 178–187 (2013).
15. Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., Fernández-Manjón, B.: Applying standards to systematize learning analytics in serious games. Comput. Stand. Interfaces. 50, 116–123 (2017).
16. eUCM: xAPI Serious Games Profile, http://w3id.org/xapi/seriousgames.
17. ADL Initiative: xAPI Specification, https://github.com/adlnet/xAPI-Spec/blob/a752217060b83a2e15dfab69f8c257cd86a888e6/xAPI.md.
18. ADL Initiative: xAPI statement generator, http://experienceapi.com/statement-generator/.
19. Perez Colado, I., Perez Colado, V., Martínez-Ortiz, I., Freire, M., Fernandez-Manjon, B.: uAdventure: The eAdventure reboot - Combining the experience of commercial gaming tools and tailored educational tools. In: IEEE Global Engineering Education Conference (EDUCON). pp. 1754–1761 (2017).
20. Perez Colado, I.: uAdventure: desarrollo del intérprete y de un emulador de videojuegos de -Adventure sobre Unity3D, http://www.e-ucm.es/drafts/e-UCM_draft_294.pdf, (2016).
21. Torrente, J., Blanco, Á., Marchiori, E.J., Moreno-ger, P., Fernández-Manjón, B., Del Blanco, Á.: <e-Adventure> Introducing Educational Games in the Learning Process. IEEE Educ. Eng. EDUCON 2010 Conf. 127, 1121–1126 (2010).
22. e-ucm: eAdventure, http://e-adventure.e-ucm.es/.
23. Elastic: Elasticsearch, https://www.elastic.co/products/elasticsearch.
24. Elastic: Kibana, https://www.elastic.co/products/kibana.
25. Docker: Docker, https://www.docker.com/.
26. OASIS: Security Assertion Markup Language (SAML), https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security.
27. IMS Global Learning Consortium: Learning Tools Interoperability, https://www.imsglobal.org/activity/learning-tools-interoperability.
28. e-ucm: RAGE Analytics GitHub Wiki Page, https://github.com/e-ucm/rage-analytics/wiki.
29. e-ucm: First Aid Game, http://first-aid-game.e-ucm.es/.
30. Pérez Colado, V.: Extendiendo uAdventure con funcionalidades de geoposicionamiento, http://www.e-ucm.es/drafts/e-UCM_draft_302.pdf, (2017).

6.2.4. Improving serious games analyzing learning analytics data: lessons learned

Full citation

Cristina Alonso-Fernández, Ivan Perez-Colado, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2018): **Improving serious games analyzing learning analytics data: lessons learned.** Games and Learning Alliance conference (GALA Conf), December 5-7, 2018, Palermo, Italy.

Abstract

Serious games adoption is increasing, although their penetration in for-mal education is still surprisingly low. To improve their outcomes and increase their adoption in this domain, we propose new ways in which serious games can leverage the information extracted from player interactions, beyond the usual post-activity analysis. We focus on the use of: (1) open data which can be shared for research purposes, (2) real-time feedback for teachers that apply games in schools, to maintain awareness and control of their classroom, and (3) once enough data is gathered, data mining to improve game design, evaluation and deployment; and allow teachers and students to benefit from enhanced feedback or stealth assessment. Having developed and tested a game learning analytics platform throughout multiple experiments, we describe the lessons that we have learnt when analyzing learning analytics data in the previous contexts to improve serious games.

Full publication

Improving Serious Games Analyzing Learning Analytics Data: Lessons Learned

Cristina Alonso-Fernández(✉), Iván Pérez-Colado, Manuel Freire, Iván Martínez-Ortiz, and Baltasar Fernández-Manjón

Facultad de Informática, Complutense University of Madrid, C/Profesor José García Santesmases 9, 28040 Madrid, Spain
{crisal03,ivanjper}@ucm.es, {manuel.freire,imartinez,balta}@fdi.ucm.es

Abstract. Serious games adoption is increasing, although their penetration in formal education is still surprisingly low. To improve their outcomes and increase their adoption in this domain, we propose new ways in which serious games can leverage the information extracted from player interactions, beyond the usual post-activity analysis. We focus on the use of: (1) open data which can be shared for research purposes, (2) real-time feedback for teachers that apply games in schools, to maintain awareness and control of their classroom, and (3) once enough data is gathered, data mining to improve game design, evaluation and deployment; and allow teachers and students to benefit from enhanced feedback or stealth assessment. Having developed and tested a game learning analytics platform throughout multiple experiments, we describe the lessons that we have learnt when analyzing learning analytics data in the previous contexts to improve serious games.

Keywords: Serious games · Learning analytics · Dashboards · Game-based learning · Stealth assessment

1 Introduction

Serious games are being successfully applied in multiple fields (e.g. military, health); however, their uptake in formal education is still poor, and usually restricted to complementary content for motivation [1]. Several reasons can explain this, including the high development cost of new games, or the difficulty for teachers to assess the acquired learning, and therefore to effectively deploy and apply games in their classes.

Moreover, very few serious games have a full formal evaluation, and those that have been evaluated are usually tested with limited numbers of users [2]. This is hardly surprising, as large-scale formal evaluations can become as expensive as creating the game. Also, the feedback from formal evaluations is often obtained too late to improve the games or their educational experience. We consider that information from In-game users interactions can benefit all phases of a serious game's lifecycle, including game design, development, piloting, acceptance, evaluation and maintenance; and should be used to improve the experience of all stakeholders involved (teachers, educators, and

M. Gentile et al. (Eds.): GALA 2018, LNCS 11385, pp. 287–296, 2019.
https://doi.org/10.1007/978-3-030-11548-7_27

students), providing each with the specific information that they need for their purposes. But this process is still too game-dependent, complex and expensive.

Analysis of in-game user interaction data has been used to improve games development in the entertainment industry, in a discipline called Game Analytics (GA). This requires data to be obtained via telemetry, and then analyzed to extract metrics, such as performance or user habits. However, the usual focus of GA is increasing user retention, playing time and revenue [3]; while serious games, particularly in education, instead seek to maximize learning or improve the learning experience.

Learning Analytics (LA) is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [4]. LA seeks to lay the groundwork to go from theory-driven to evidence-based education, where data can be used to improve educational scenarios [5]. This approach can be extended to serious games, where in-game user interaction data can benefit their creation and applicability in real environments, in a discipline that we call Game Learning Analytics (GLA) [6].

In this paper, we wish to go beyond the usual post-game session analysis, and focus instead on three scenarios where GLA can be especially helpful. First, the data extracted, if done in a systematic and standardized way, can be not only used for improving the games but also openly shared for research purposes. Second, in the context of applying games in education, all stakeholders involved could benefit from information from the actions taken in the game, directly during the session. Finally, after sufficient data has been gathered, deeper analysis can be helpful to obtain richer information for all stakeholders, and inform improvements in several stages of the game's lifecycle.

2 Obtaining In-Game User Interaction Data

The first step to gain insights from in-game user interactions is to ensure that all data with the potential to yield such insights is adequately collected. Experimental design and deployment should comply with all the legal regulations (e.g. users' consent). We consider three main pillars that data management must ensure:

- **Anonymization:** when possible, data must be adequately anonymized so no personal details are attached to the student data (e.g. using randomly generated codes). This will help to comply with regulations on data privacy [7].
- **Collection:** data collection must be non-intrusive and transparent, to avoid interrupting the students' gameplay. Collection can be greatly improved using a standard tracking model that simplifies and standardizes this process.
- **Storage:** data received from games should be collected in a server that can efficiently manage large amounts of data in a secure way. If data is collected in a specific format, the storage system should also be prepared to validate and handle that format.

Our research group has developed a GLA System that is currently been improved and extended as part of two EU H2020 projects (RAGE and BEACONING). With this analytics platform, we have already conducted several experiments that follow the

above data-management guidelines, collecting data from thousands of game sessions. Some of the results and conclusions drawn from these experiences are detailed in the following sections, since we have used the resulting data for each of the three applications described in this paper: research, real-time reports, and deeper offline analysis.

2.1 Standardizing Data Collection: Experience API Serious Games Profile

To systematize and standardize data collection we propose the use of the Experience API Serious Games Profile (xAPI-SG for short), described in detail in [8].

As previously mentioned, it is mandatory to comply with all personal data privacy regulations, capturing only the relevant data and using anonymization whenever possible before storage, so no data can be traced back to specific students. For analytics, pseudo-anonymization techniques can be used, where the manager of a session assigns random tokens to players that use them to access the game. The tokens tie all data received from each player together, while providing no information of their identity. When required, teachers can retain the correspondence between anonymous tokens and students that use them; in such cases, this link must be managed outside the game and the analytics system. Additionally, best practices require informed consent forms disclosing both the intended experimental design and how collected data will be used.

Servers that can store xAPI data are usually called Learning Record Stores (LRS), and generally allow limited query capabilities. Every server and technology used in the tracking architecture should be ready to deal with large amounts of data (*big data*) as the number of traces generated by a single player may be large; and if the system is successful, large amounts of users generating many interactions per second can easily overwhelm low-capacity solutions. To ensure scalability, multiple servers that can share the load are an obvious choice; however, this also increases the chances of at least one of them failing or becoming unreachable, forcing truly scalable analytics implementations to be distributed, redundant, and fault-resistant.

Data tracked from serious games using an open format such as xAPI-SG, when suitably anonymized, can be easily shared with other researchers. Open sharing of research data and publications are among the tenets of the Open Science movement, with initiatives such as the European Commission's OpenAIRE [9] or CERN's Zenodo [10], which seek to ensure open access of research data and publications, respectively.

3 Uses of Analytics Data to Improve Serious Games

Users of serious games can benefit from collected data at several stages: (1) at real time, to provide real-time feedback to teachers and students, (2) after the session is finished, to provide detailed feedback, and (3) after sufficient data has been collected, through enriched feedback based on data mining. In this paper we are mostly interested in (1) and (3), since (2) is generally known and explored in many other resources.

Figure 1, has been adapted from the Learning Analytics Framework (LAF) described in [11], to add and highlight the use of open data, real-time feedback, and data-mining. The LAF did not envision serious games as sources for analytics data, and

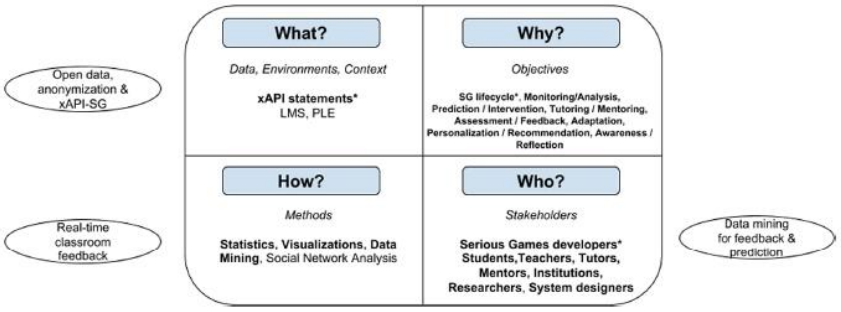

Fig. 1. Contributions of this paper (ovals), and an adapted version of the Learning Analytics framework, as described in [11]. Elements with an asterisk were not present in the original framework. Bold-face text highlights framework elements affected by our contributions.

predates the appearance of xAPI. As seen in the figure, the contributions described in this work have consequences for most if not all elements considered in the LAF. For each of the 4 questions considered in the LAF:

- **What** kind of data does the system gather, manage and use for analysis? While the LAF mentions e-learning systems as sources, we propose the use of xAPI anonymized statements from games.
- **Why** does the system analyze the collected data? We extend the goals of the LAF to improve serious games, from design to development, deployment, and maintenance. Real-time feedback is key for monitoring and classroom intervention, while data mining allows enhanced feedback once enough data has been gathered.
- **How** does the system perform the analysis of the collected data? We apply most of the methods envisioned by the LAF, even though our current focus is on single-player games.
- **Who** is targeted by the analysis? Stakeholders now include the actual game developers, in addition to the learners and teachers that gain, among others, feedback and assessment, or educational institutions that wish to know the outcomes of applying games in education. Researchers can also benefit from open research data.

4 Game Learning Analytics Real-Time Applications

Different stakeholders can benefit from (near) real-time feedback. In this section, we adopt the common scenario of using serious games in education as part of a lesson in a classroom environment, focusing on real-time feedback for teachers and students.

Real-time feedback is available as soon as the game starts to be played; to allow teachers to monitor and perform timely interventions, easy-to-understand feedback must be quickly generated. For example, a student that stops playing can trigger an alert that allows the teacher to walk over to find out the cause; or a student that is advancing much quicker than the rest of the class may benefit from the teacher suggesting additional tasks to attempt. Such simple scenarios illustrate real-time applications where the information

collected from interaction data can help teachers and students. Certain types of visual analytics are particularly suited for real-time feedback, especially when combined in dashboards. The ideal content of these dashboards will depend on the game, delivery environment, and the metrics and KPIs that are most relevant to each stakeholder.

4.1 Real-Time Information for Teachers

For teachers, visual analytics provide an easy way to explore the information gathered from their students' interactions. Analytics dashboards present aggregations of individual visualizations, each providing insight into specific aspects, such as progress, errors, or choices taken. Visualizations can also display actionable feedback to locate students that get stuck, or suggest additional work that may interest advanced students.

We have conducted multiple experiments with students to test our data gathering, real-time analytics and dashboards; the latest, as of this writing, with over 1000 students, seeking to validate a serious game that raises awareness on cyberbullying [12]. Previous experiments include games that teach first aid techniques [13], or that were geared towards cognitively impaired users (e.g. with Down Syndrome or Autism) [14], where dashboards were the only option to follow the progress of players.

Figure 2 describes some of the visualizations included in the teacher dashboard used in the latest experiments to provide real-time feedback in classroom settings [15]. The dashboard uses xAPI-SG concepts such as *completables* (e.g. levels) and *alternatives* (e.g. multiple-choice questions). The visualizations depicted inform users on (a) *correct and incorrect alternatives selected*: the number of correct and incorrect answers selected as *alternatives* for each player, and therefore the general knowledge of players; (b) *total session players*: the number of students that have started the game; (c) *maximum progress of players per completable:* for each *completable*, the progress achieved by each player, and therefore whether students are finishing or struggling to continue; and finally (d) *games started and completed*: a pie-chart that displays the number of games that have been started and completed, providing an overview of the students that have started and finished; and indirectly, how many students are still playing.

These visualizations aim to provide general information from gameplays (e.g. progress, answers) to teachers, allowing them to understand it with minimal effort. Data can be used to trigger alerts or warnings in specific situations that require immediate action for teachers (e.g. a player has been inactive for too long). Figure 3 shows the general view of alerts and warnings; clicking on a specific student, teachers can see details of alerts and warnings triggered by that student's actions, and act accordingly.

Improvements are being considered for the visualizations in Fig. 2, based on the feedback collected from teachers. For example, simplifying the visualizations by adding clearer titles and legends, and showing general metrics that provide a quicker overview of the most critical information (e.g. questions failed most, critical areas in-game). We have also determined that providing additionally recommended actions is well-received by teachers (e.g. specific student needs help). These recommendations can help teachers to improve their classes, linking the information provided by LA with actions to support students learning [16]. For instance, teachers requested reports on the topics with the highest error ratios, to allow them to be reviewed before any others.

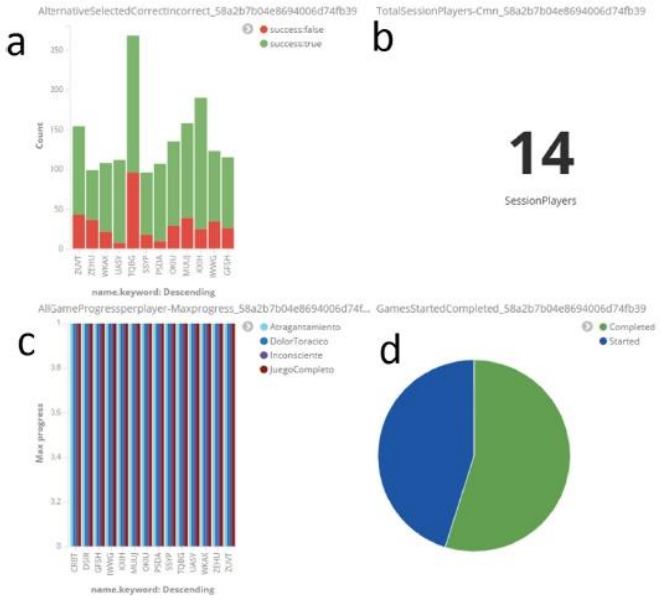

Fig. 2. Some of the visualizations included on the default teacher dashboard.

Fig. 3. Alerts and warnings general view (top part). Clicking on a specific student displays detailed information on the warnings and alerts triggered by that student (bottom part).

4.2 Real-Time Information for Students

Student dashboards provide information on performance and in-game outcomes, allowing them to easily assess their strengths and weakness. Current solutions for learners' dashboards present several issues that should be considered. It is common to compare the results of students with their class or with average results (e.g. scores, times-to-finish) from their classmates; however, some researches have pointed out that this may demotivate those students who do not reach at least average rankings [17]. Authors of [18] concluded that most educational concepts used to design LA

dashboards focus on self-regulated training by displaying their own data to players. These dashboards generally fail to use awareness and reflection to improve competencies (e.g. cognitive, behavioral or emotional) and usually promote competition instead of knowledge mastery.

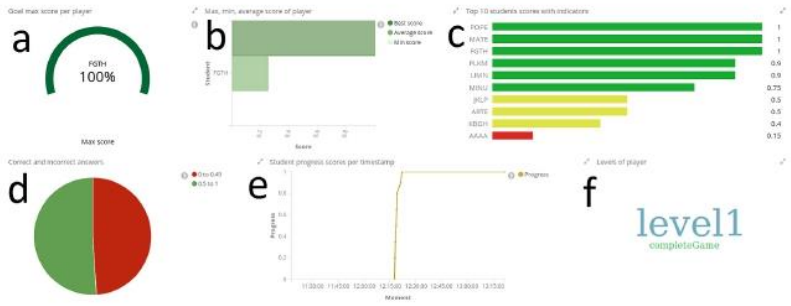

Fig. 4. Sample student dashboard showing information about scores, errors or progress.

Figure 4 provides a sample dashboard with typical information shown to students: maximum score achieved (visualization labelled *a*); maximum, average and minimum scores of the player (*b*); a comparison with other players in a leaderboard (*c*), which uses traffic-light colors to display ranges of players scores; correct and incorrect answers (*d*); student progress over time (*e*); and levels completed (*f*).

Another area where data can be exploited at real-time to benefit students is that of adaptive learning experiences. Games can adapt their difficulty in real-time in response to players' in-game performance. The authors of [19] described the results of experiments comparing adaptive, non-adaptive games and other non-adaptive learning activities, concluding that, although all activities reached equal levels of motivation, the adaptive game resulted in significantly higher learning outcomes.

5 Offline Data Analysis

Apart from displaying real-time information, data collected during several gameplays can be further analyzed to yield additional insights. Data mining processes can be applied to extract patterns of use, which can be leveraged to improve the game for future deployments. Educational institutions and higher-education administrations can also benefit from aggregated data that can quantify the extent to which the use of games in class benefits learning, allowing them to make evidence-based decisions on the value of using serious games in their classrooms.

Game developers and designers can obtain feedback from actual classroom gameplays to improve both the game and the learning design. For instance, they may find errors in the game, unreachable areas, levels that are too difficult or too easy for players, a more precise determination of average playing time, etc. In this sense, analysis can certainly help to improve the iterative design for subsequent versions of the game.

Data gathered can also be used to categorize players, creating different profiles for targeted feedback, a process which can be automated using data mining (considering some limits such as data complexity or the algorithms' efficiency). This feedback may include hints to help students or even changes in level difficulty to avoid a decrease in motivation [20]. Clusters of players that show different behaviors and characteristics may provide clues on how different learning elements affect each type of players [21]. In one experiment, we collected data for more than 200 students playing a serious game that teaches first aid techniques. Using data mining, we managed to classify players based on their actions and results to identify clusters of player profiles, and even the discovery of those in-game actions that had greater influence on player outcomes [13].

One of the latest steps we have carried out to improve the lifecycle of serious games using GLA data focuses on improving evaluation methods. So far, serious games are commonly evaluated using costly pre-post experiments [2]. We consider that their application in education would benefit from a quicker and cheaper evaluation process. To this end, we have proposed and tested the use of data mining to predict the results of the pre- and post- tests based on interaction data; while only possible once a sufficiently-large training set has been collected, the technique avoids the need of conducting tests – at least for players that are similar to those that the system was trained with.

This can be considered as another step in stealth assessment. Serious games can also become powerful assessment tools, even though the exact characteristics of games that best allow these assessments is still not entirely clear. Stealth assessment [22] is the practice of embedding assessment in a gaming environment in a non-intrusive way, analyzing gameplay actions to infer exactly what players know at each point in time; it is, in this sense, and extension of the evaluation without pre- and post-test described in the previous paragraph. For this discipline, it is still important to improve the serious games themselves, ensuring that their application is effective and making assessment more valid and reliable.

6 Conclusions

In-game users' interaction data from serious games can be exploited to provide a wide variety of insight on the educational process of different stakeholders. Developers can use it to improve the full lifecycle of games. Teachers can gain real-time insights of student behavior, allowing them to help students playing, or to summarize a session when discussing it with students once they have finished. Students can get feedback on their performance, including their strengths and weaknesses. Researchers can benefit from open access to shared research data. Educators can obtain metrics on the efficacy of games application on their institutions.

When collecting data, many issues need to be addressed, including anonymization, which is especially relevant when working with minors or to allow collected data to be openly shared for research purposes.

In real-time scenarios, visualizations, alerts and warnings can help teachers to gain insights of the whole classroom, while students can track their performance and

compare it to that of their peers; both uses of data provide information that allows teachers and students to make better decisions while the games are still in play.

After data has been collected, data mining techniques can provide further information to improve game design, deployment and evaluation. To improve evaluation, the latest line of work continues with experiments that follow the usual evaluation structure (pre-test, gameplay, post-test) and track interaction data to later predict previous and subsequent knowledge; and compare those predictions against actual data collected in the tests.

Adoption of serious games in schools could greatly improve with the feedback retrieved from interaction data, ideally with a general game learning analytics system that standardizes data tracking, collection, analysis and visualization, extracting useful information to be given back to the stakeholders involved. Whichever the approach, we consider that data-driven solutions that take advantage of the power of game learning analytics are essential to guide the future application of serious games.

Acknowledgments. This work has been partially funded by Regional Government of Madrid (eMadrid S2013/ICE-2715), by the Ministry of Education (TIN2017-89238-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676, Erasmus+IMPRESS 2017-1-NL01-KA203-035259).

References

1. Popescu, M., et al.: Serious games in formal education: discussing some critical aspects. In: Proceedings of the 5th European Conference on Games Based Learning, pp. 486–493 (2011)
2. Calderón, A., Ruiz, M.: A systematic literature review on serious games evaluation: an application to software project management. Comput. Educ. **87**, 396–422 (2015)
3. El-Nasr, M., Drachen, A., Canossa, A.: Game Analytics: Maximizing the Value of Player Data. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4769-5
4. Long, P., Siemens, G.: Penetrating the fog: analytics in learning and education. Educ. Rev. **46**, 31–40 (2011)
5. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief, pp. 1–57. SRI International, Washington, DC (2012)
6. Freire, M., Serrano-Laguna, Á., Iglesias, B.M., Martínez-Ortiz, I., Moreno-Ger, P., Fernández-Manjón, B.: Game learning analytics: learning analytics for serious games. In: Spector, M., Lockee, B., Childress, M. (eds.) Learning, Design, and Technology, pp. 1–29. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-17727-4_21-1
7. European Union: Regulation (EU) 2016/679. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679. Accessed Sept 2018
8. Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., Fernández-Manjón, B.: Applying standards to systematize learning analytics in serious games. Comput. Stand. Interfaces **50**, 116–123 (2017)
9. European Commission: OpenAIRE. https://www.openaire.eu/. Accessed Sept 2018
10. CERN, OpenAIRE, EC: Zenodo. https://zenodo.org/. Accessed Sept 2018
11. Chatti, M.A., et al.: Learning Analytics: Challenges and Future Research Directions. E-Learning Education (2015)

12. Morata, A.C.: Videojuegos Como Herramienta Educativa En La Escuela: Concienciando Sobre El Ciberbullying (Master Thesis) (2017)
13. Alonso-Fernández, C.: Applying data mining techniques to game learning analytics (Master Thesis) (2017)
14. Cano, A.R., Fernández-Manjón, B., García-Tejedor, Á.J.: Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. Br. J. Educ. Technol. **49**, 659 (2018)
15. Alonso-Fernandez, C., Calvo, A., Freire, M., Martinez-Ortiz, I., Fernandez-Manjon, B.: Systematizing game learning analytics for serious games. In: IEEE Global Engineering Education Conference (EDUCON). IEEE (2017)
16. Bakharia, A., et al.: A conceptual framework linking learning design with learning analytics. In: Proceedings of the 6th International Conference on Analytics and Knowledge 2016, LAK16, pp. 329–338 (2016)
17. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: learning analytics are about learning. TechTrends **59**, 64–71 (2015)
18. Jivet, I., Scheffel, M., Drachsler, H., Specht, M.: Awareness is not enough: pitfalls of learning analytics dashboards in the educational practice. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 82–96. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66610-5_7
19. Sampayo-Vargas, S., Cope, C.J., He, Z., Byrne, G.J.: The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. Comput. Educ. **69**, 452–462 (2013)
20. Shute, V., Ke, F., Wang, L.: Assessment and adaptation in games. In: Wouters, P., van Oostendorp, H. (eds.) Instructional Techniques to Facilitate Learning and Motivation of Serious Games. AGL, pp. 59–78. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-39298-1_4
21. Loh, C.S., Sheng, Y., Ifenthaler, D.: Serious Games Analytics. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-05834-4
22. Shute, V.J., Moore, G.R.: Consistency and validity in game-based stealth assessment. In: Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective (2017)

6.2.5. Applications of learning analytics to assess serious games

Full citation

Cristina Alonso-Fernández, Ana Rus Cano, Antonio Calvo-Morata, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón (2018): **Applications of learning analytics to assess serious games**. 2nd Annual Learning & Student Analytics Conference (LSAC), October 22-23, 2018, Amsterdam, The Netherlands.

Abstract

We summarize our experiences regarding three applications of Learning Analytics (LA) for Serious Games (SGs) with different purposes: A. Validate and deploy games in schools. The SG Conectado has been designed to address social problems (bullying and cyberbullying). B. Validate game design when information cannot be directly gathered from users. The SG Downtown was designed for improving independent life of users with Intellectual Disabilities (ID) who struggle with communication issues. C. Improve evaluation and deployment of games. The SG First Aid Game was already validated and data mining models were applied to predict knowledge after playing. All three games have been tested with target users in actual classrooms, as described in the following section. Results and implications of the use of analytics in those three scenarios are later explained.

Full publication

LSAC 2018

Applications of Learning Analytics to assess Serious Games

Cristina Alonso-Fernández, Ana Rus Cano, Antonio Calvo-Morata,
Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón
Facultad de Informática, Complutense University of Madrid

Purpose

We summarize our experiences regarding three applications of Learning Analytics (LA) for Serious Games (SGs) [1] with different purposes:

A. Validate and deploy games in schools. The SG ***Conectado*** has been designed to address social problems (bullying and cyberbullying) [2].
B. Validate game design when information cannot be directly gathered from users. The SG ***Downtown*** was designed for improving independent life of users with Intellectual Disabilities (ID) who struggle with communication issues [3].
C. Improve evaluation and deployment of games. The SG ***First Aid Game*** was already validated and data mining models were applied to predict knowledge after playing [4].

All three games have been tested with target users in actual classrooms, as described in the following section. Results and implications of the use of analytics in those three scenarios are later explained.

Design

A. Conectado

Conectado is a videogame to raise bullying and cyberbullying awareness for students (12-17 years old). The game places the player in the role of a student suffering cyberbullying by schoolmates and promotes empathy through mini-games presented as nightmares and dialogues with other characters. Choices taken in the game alter the story, e.g. the ending is determined by the relationship with classmates and parents and w9hether players have asked the teacher for help or not.

Figure 1. Captures of *Conectado*.

The game was tested with N=257 high-school students from three educational centers in Spain, in June 2017.

B. DownTown

Downtown, a Subway Adventure is a spy game for players with ID (18-45 years old) to train them in using the subway transportation system of Madrid (Spain). The game was developed in a 3D realistic perspective so players can identify the scenarios with reality. Users can navigate in the game as in real life, choosing the routes from one station to another. *Downtown* also includes quests to train daily skills (e.g. independence, long and short memory, spatial vision) and social aspects to promote users' independent life.

Figure 2. Captures of *DownTown*.

The game was tested with N=51 adults with ID (Down Syndrome, mild cognitive disability or certain types of Autism Spectrum Disorder) in May-June 2017. Students played a total of 3 hours. Interaction data captured included character and accessibility preferences, timestamps, attempts, correct/incorrect stations during the route, number of clicks in the interface and progress.

C. First Aid Game

First Aid Game aims to instruct cardiopulmonary resuscitation (CPR) maneuvers for students (12-16 years old). The knowledge to be learned is assessed with questions that players can retry when choosing a wrong option. The more mistakes the user makes, the less score is given for that situation.

The game was validated in a usual pre-post experiment with a control group in 2011 with more than 300 students in four secondary schools of Aragon (Spain) [5]. Learning was slightly lower in the experimental group but still significant.

Figure 3. Captures of *First Aid Game*.

The game was tested with N=227 high-school students from one school in Madrid (Spain) in January-February 2017. Data captured from tests included knowledge in pre- and post- tests, opinion; interaction data comprised scores, correct/incorrect answers, interactions and times.

Results

A. Conectado

The increase in cyberbullying awareness was measured in pre-post tests, each containing eighteen 7-point Likert items to validate the game. The average score in pre-test was 5.72 (SD=1.26), compared to 6.38 (SD=1.11) in post-test, a statistically-significant effect.

B. DownTown

Most students (85.8%) reached a destination; half of the mistakes (50.8%) occurred during the first 30 minutes of playing (once students completed a few routes to understand the mechanics).

C. First Aid Game

Best models for predicting pass/fail obtained 89% precision, 98% recall and 10% misclassification rate; for predicting scores in range [1-15], mean error was 1.5 (SD=1.33). Predictions without pre-test information were slightly worse but differences were not significant.

Implications

We have summarized our experiences and results for three different applications of LA for SGs:

A. Validate and deploy a SG in school that increases cyberbullying awareness.
B. Validate a SG that trains using the subway without explicit feedback from students.
C. Improve evaluation and deployment of a SG predicting knowledge after playing to avoid carrying out the post-test.

All three applications are intended to foster the use of Serious Games in real contexts for different targets, simplifying their validation, evaluation and deployment using Game Learning Analytics data.

Acknowledgments

This work has been partially funded by Regional Government of Madrid (eMadrid S2013/ICE-2715), by the Ministry of Education (TIN2017-89238-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676, Erasmus+ IMPRESS 2017-1-NL01-KA203-035259).

Resources

[1] M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, “Game Learning Analytics: Learning Analytics for Serious Games,” in Learning, Design, and Technology, Cham: Springer International Publishing, 2016, pp. 1–29.
[2] Calvo-Morata, A., Rotaru, D.C., Alonso-Fernández, C., Freire-Morán, M., Martínez-Ortiz, I., Fernández-Manjón, B. (2018). Validation of a Cyberbullying Serious Game Using Game Analytics. IEEE Transactions on Learning Technologies (*under review*).

[3] Cano, A. R., Fernández-Manjón, B., & García-Tejedor, Á. J. (2018). Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. British Journal of Educational Technology. https://doi.org/10.1111/bjet.12632

[4] Alonso-Fernandez, C., Martinez-Ortiz, I., Caballero Roldán, R., Freire, M., Fernández-Manjón, B. (2018) Improving serious games evaluation using data mining techniques. IEEE Transactions on Learning Technologies (*submitted*).

[5] Marchiori, E.J., Ferrer, G., Fernández-Manjón, B., Povar Marco, J., Suberviola González, J.F., Giménez Valverde, A. (2012). Video-game instruction in basic life support maneuvers. Emergencias. 24:433-7.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adamo-Villani, N., Haley-Hermiz, T., & Cutler, R. (2013). Using a Serious Game Approach to Teach "Operator Precedence" to Introductory Programming Students. In *2013 17th International Conference on Information Visualisation* (pp. 523–526). IEEE. https://doi.org/10.1109/IV.2013.70

ADL. (2012). Experience API. Retrieved March 20, 2016, from https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/

ADL. (2017). xAPI Profiles. Retrieved from https://adlnet.github.io/xapi-profiles/

Agarwal, S. (2014). *Data mining: Data mining concepts and techniques*. *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2017). Systematizing game learning analytics for serious games. In *2017 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1111–1118). IEEE. https://doi.org/10.1109/EDUCON.2017.7942988

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2020a). Evidence-based evaluation of a serious game to increase bullying awareness. *Interactive Learning Environments*, 1–11. https://doi.org/10.1080/10494820.2020.1799031

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2020b). Simplifying the Validation and Application of Games with Simva. In *Emerging Technologies for Education* (pp. 337–346). https://doi.org/10.1007/978-3-030-38778-5_37

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301–309. https://doi.org/10.1016/j.chb.2019.05.036

Alonso-Fernández, C., Pérez-Colado, I., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Improving Serious Games Analyzing Learning Analytics Data: Lessons Learned. In *Games and Learning Alliance: 7th International Conference, GALA 2018, Palermo, Italy, December 5–7, 2018, Proceedings* (Vol. 10653, pp. 287–296). https://doi.org/10.1007/978-3-030-11548-7_27

Alonso-Fernández, C., Perez-Colado, I. J., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2019). Using Simva to evaluate serious games and collect game learning analytics data. In *LASI Spain 2019: Learning Analytics in Higher Education* (pp. 22–34). Retrieved from https://pubman.e-ucm.es/drafts/e-UCM_draft_343.pdf

Alonso-Fernandez, C., Perez-Colado, I. J., Calvo-Morata, A., Freire, M., Ortiz, I. M., & Manjon, B. F. (2020). Applications of Simva to Simplify Serious Games Validation and Deployment. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *15*(3), 161–170. https://doi.org/10.1109/RITA.2020.3008117

Alonso-Fernández, C., Rotaru, D. C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2017). Full Lifecycle Architecture for Serious Games: Integrating Game Learning Analytics and a Game Authoring Tool. In *Lecture Notes in Computer Science* (Vol. 10622 LNCS, pp. 73–84). https://doi.org/10.1007/978-3-319-70111-0_7

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2021). Data science meets standardized game learning analytics. In *2021 IEEE Global Engineering Education Conference (EDUCON)*.

Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2021). Improving evidence-based assessment of players using serious games. *Telematics and Informatics*.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, *36*(3), 350–358. https://doi.org/10.1111/jcal.12405

Alonso-Fernández, C., Rus Cano, A., Calvo-Morata, A., Freire, M., Martínez-Ortiz, & Fernández-Manjón, B. (2018). Applications of Learning Analytics to assess Serious Games. In *2nd Annual Learning & Student Analytics Conference (LSAC)*. Amsterdam.

Álvarez-García, D., Núñez Pérez, J. C., & Dobarro González, A. (2013). Cuestionarios para evaluar la violencia escolar en Educación Primaria y en Educación Secundaria: CUVE3-EP y CUVE3-ESO. *Apuntes de Psicología*, *31*(2), 191–202. Retrieved from

http://www.apuntesdepsicologia.es/index.php/revista/article/view/322/296

Asociación Servicio Interdisciplinar de Atención a las Drogodependencias (SIAD). (2014). Aislados. Retrieved November 13, 2016, from http://www.aislados.es/zona-educadores/

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning*, *32*(3), 267–280. https://doi.org/10.1111/jcal.12128

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Washington, DC: SRI International*, 1–57. Retrieved from https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Calderón, A., & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, *87*, 396–422. https://doi.org/10.1016/j.compedu.2015.07.011

Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2020). Creating awareness on bullying and cyberbullying among young people: validating the effectiveness and design of the serious game Conectado (submitted). *Telematics and Informatics*.

Calvo-Morata, A., Rotaru, D. C., Alonso-Fernandez, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2020). Validation of a Cyberbullying Serious Game Using Game Analytics. *IEEE Transactions on Learning Technologies*, *13*(1), 186–197. https://doi.org/10.1109/TLT.2018.2879354

Calvo Morata, A. (2020). *Uso de técnicas de learning analytics para la validación, mejora y aplicación de juegos serios en la clase aplicado al ciberbullying*. Universidad Complutense de Madrid.

Cano, A. R., Fernández-Manjón, B., & García-Tejedor, Á. J. (2018). Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. *British Journal of Educational Technology*, *49*(4), 659–672. https://doi.org/10.1111/bjet.12632

Center for Game Science at the University of Washington. (2016). Treefrog Treasure. Retrieved November 15, 2016, from http://centerforgamescience.org/blog/portfolio/treefrog-treasure/

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5/6), 318. https://doi.org/10.1504/IJTEL.2012.051815

Chaudy, Y., Connolly, T., & Hainey, T. (2014). Learning Analytics in Serious Games: a Review of the Literature. *Ecaet 2014*, (March 2016).

Cheng, M.-T., Rosenheck, L., Lin, C.-Y., & Klopfer, E. (2017). Analyzing gameplay data to inform feedback loops in The Radix Endeavor. *Computers & Education*, *111*, 60–73. https://doi.org/10.1016/j.compedu.2017.03.015

Chung, G. K. W. K. (2015). Guidelines for the Design and Implementation of Game Telemetry for Serious Games Analytics. In *Serious Games Analytics* (pp. 59–79). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_3

Clark, D. B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., & Sengupta, P. (2012). Driving Assessment of Students' Explanations in Game Dialog Using Computer-Adaptive Testing and Hidden Markov Modeling. In *Assessment in Game-Based Learning* (pp. 173–199). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3546-4_10

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*(2), 661–686. https://doi.org/10.1016/j.compedu.2012.03.004

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2017). Assessing Whether Students Seek Constructive Criticism: The Design of an Automated Feedback System for a Graphic Design Task. *International Journal of Artificial Intelligence in Education*, *27*(3), 419–447. https://doi.org/10.1007/s40593-016-0137-5

de Klerk, S., & Kato, P. (2017). The Future Value of Serious Games for Assessment: Where Do We Go Now?. *Journal of Applied Testing Technology*, *18*(February), 32–37.

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., … Lester, J. C. (2018). Detecting and Addressing Frustration in a Serious Game for Military Training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193. https://doi.org/10.1007/s40593-017-0152-1

Denden, M., Tlili, A., Essalmi, F., & Jemni, M. (2018). Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors. *Smart Learning Environments*, *5*(1), 1–19. https://doi.org/10.1186/s40561-018-0078-6

Dicerbo, K. E. (2013). Game-based assessment of persistence. *Educational Technology and Society*, *17*(1), 17–28.

DiCerbo, K. E., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, G. T. (2015). An Application of Exploratory Data Analysis in the Development of Game-Based Assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics* (pp. 319–342). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_14

Dörner, R., Göbel, S., Effelsberg, W., & Wiemeyer, J. (Eds.). (2016). *Serious Games*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40612-1

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*(x), 155–161. https://doi.org/10.1.1.10.4845

Elaachak, L., Belahbibe, A., & Bouhorma, M. (2015). Towards a System of Guidance, Assistance and Learning Analytics Based on Multi Agent System Applied on Serious Games. *International Journal of Electrical and Computer Engineering (IJECE) Journal*, *5*(2), 2088–8708. Retrieved from http://iaesjournal.com/online/index.php/IJECE

ElAtia, S., Ipperciel, D., & Zaïane, O. R. (2016). *Data Mining and Learning Analytics*. (S. ElAtia, D. Ipperciel, & O. R. Zaïane, Eds.), *Data Mining And Learning Analytics: Applications in Educational Research*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118998205

Electronic Arts. (2013). SimCityEDU: Pollution Challenge! Retrieved December 18, 2020, from http://www.simcityedu.org/

Electronic Arts Games. (2019). SimCity BuildIt. Retrieved December 18, 2020, from https://www.ea.com/es-es/games/simcity/simcity-buildit

European Commission. (2018). 2018 reform of EU data protection rules. Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Evans, K. H., Daines, W., Tsui, J., Strehlow, M., Maggio, P., & Shieh, L. (2015). Septris. *Academic Medicine*, *90*(2), 180–184. https://doi.org/10.1097/ACM.0000000000000611

Firaxis Games. (2016). Civilization. Retrieved December 18, 2020, from https://civilization.com

Forsyth, C., Pavlik, P., Graesser, A., Cai, Z., Germany, M.-L., Millis, K., … Halpern,

D. (2012). Learning Gains for Core Concepts in a Serious Game on Scientific Reasoning. *Proceedings of the 5th International Conference on Educational Data Mining*, 1–4. Retrieved from http://w.optimallearning.org/people/Articles/edm2012_short_2.pdf

Frederick-Recascino, C., Liu, D., Doherty, S., Kring, J., & Liskey, D. (2013). Articulating an Experimental Model for the Study of Game-Based Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8018 LNCS, pp. 25–32). https://doi.org/10.1007/978-3-642-39226-9_4

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game Learning Analytics: Learning Analytics for Serious Games. In *Learning, Design, and Technology* (pp. 1–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_21-1

Freitas, S. de, & Gibson, D. (2014). Exploratory learning analytics methods from three case studies. *Rhetoric and Reality: Critical Perspectives on Educational Technology. Proceedings of Ascilite Dunedin 2014*, 383–388.

Garaigordobil, M., & Aliri, J. (2013). Ciberacoso ("cyberbullying") en el País Vasco: Diferencias de sexo en víctimas, agresores y observadores. *Behavioral Psychology/ Psicologia Conductual*, *21*(3), 461–474.

García-Tejedor, Á. J., Cano, A. R., & Fernández-Manjón, B. (2016). GLAID: Designing a Game Learning Analytics Model to Analyze the Learning Process in Users with Intellectual Disabilities. In *6th EAI International Conference on Serious Games, Interaction and Simulation*. Porto, Portugal. Retrieved from http://sgamesconf.org/2016/show/technical-session

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Ghergulescu, I., & Muntean, C. H. (2016). ToTCompute: A Novel EEG-Based TimeOnTask Threshold Computation Mechanism for Engagement Modelling and Monitoring. *International Journal of Artificial Intelligence in Education*, *26*(3), 821–854. https://doi.org/10.1007/s40593-016-0111-2

Gibson, D., & Clarke-Midura, J. (2015). Some Psychometric and Design Implications of Game-Based Learning Analytics. *E-Learning Systems, Environments and Approaches*, (CELDA), 247–261. https://doi.org/10.1007/978-3-319-05825-2_17

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer*

Assisted Learning, *29*(3), 207–219. https://doi.org/10.1111/j.1365-2729.2012.00489.x

GTLHistory. (2020). Games to learn history. Retrieved October 22, 2020, from https://www.gtlhistory.com/

Gweon, G.-H., Lee, H.-S., Dorsey, C., Tinker, R., Finzer, W., & Damelin, D. (2015). Tracking student progress in a game-like learning environment with a Monte Carlo Bayesian knowledge tracing model. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (pp. 166–170). New York, New York, USA: ACM Press. https://doi.org/10.1145/2723576.2723608

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., & Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education*, *102*, 202–223. https://doi.org/10.1016/j.compedu.2016.09.001

Halverson, R., & Owen, V. E. (2014). Game-based assessment: an integrated model for capturing evidence of learning in play. *International Journal of Learning Technology*, *9*(2), 111. https://doi.org/10.1504/ijlt.2014.064489

Han, J., Kamber, M., & Pei, J. (2012). Introduction. In *Data Mining* (pp. 1–38). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2015). Replay Analysis in Open-Ended Educational Games. In *Serious Games Analytics* (pp. 381–399). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_17

Hauge, J. B., Berta, R., Fiucci, G., Manjon, B. F., Padron-Napoles, C., Westra, W., & Nadolski, R. (2014). Implications of Learning Analytics for Serious Game Design. In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 230–232). IEEE. https://doi.org/10.1109/ICALT.2014.73

Heeter, C., Lee, Y.-H., Medler, B., & Magerko, B. (2013). Conceptually Meaningful Metrics: Inferring Optimal Challenge and Mindset from Gameplay. In *Game Analytics* (pp. 731–762). London: Springer London. https://doi.org/10.1007/978-1-4471-4769-5_32

Hernández-Lara, A. B., Perera-Lluna, A., & Serradell-López, E. (2019). Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior*, *92*, 600–612. https://doi.org/10.1016/j.chb.2018.03.001

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016).

Using game analytics to evaluate puzzle design and level progression in a serious game. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 440–448). New York, New York, USA: ACM Press. https://doi.org/10.1145/2883851.2883953

Horn, B., Hoover, A. K., Barnes, J., Folajimi, Y., Smith, G., & Harteveld, C. (2016). Opening the Black Box of Play. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16* (pp. 142–153). New York, New York, USA: ACM Press. https://doi.org/10.1145/2967934.2968109

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2016). *A Practical Guide to Support Vector Classification*. Taipei. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Iglesias, B. M., Fernandez-Vara, C., & Fernandez-Manjon, B. (2013). E-Learning Takes the Stage: From La Dama Boba to a Serious Game. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *8*(4), 197–204. https://doi.org/10.1109/RITA.2013.2285023

interFUEL, L. (2006). Darfur is Dying. Retrieved from http://www.gamesforchange.org/play/darfur-is-dying/

Irizarry, R. A. (2019). *Introduction to Data Science*. *Introduction to Data Science*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429341830

Jaccard, D., Hulaas, J., & Dumont, A. (2017). *Using Comparative Behavior Analysis to Improve the Impact of Serious Games on Students' Learning Experience*. (J. Dias, P. A. Santos, & R. C. Veltkamp, Eds.) (Vol. 10653). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-71940-5

Jupyter Team. (2020). Jupyter Projects. Retrieved November 1, 2020, from https://jupyter.readthedocs.io/en/latest/projects/content-projects.html

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior*, *72*, 757–770. https://doi.org/10.1016/j.chb.2016.09.062

Käser, T., Busetto, A. G., Solenthaler, B., Baschera, G. M., Kohn, J., Kucian, K., … Gross, M. (2013). Modelling and optimizing mathematics learning in children. *International Journal of Artificial Intelligence in Education*, *23*(1–4), 115–135. https://doi.org/10.1007/s40593-013-0003-7

Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017). Modeling exploration strategies to predict student performance within a learning environment and beyond. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17* (pp. 31–40). New York, New York, USA: ACM Press.

https://doi.org/10.1145/3027385.3027422

Kato, P. M., & Klerk, S. De. (2017). Serious Games for Assessment: Welcome to the Jungle. *Journal of Applied Testing Technology*, *18*, 1–6.

Ke, F., Shute, V., Clark, K. M., & Erlebacher, G. (2019). *Interdisciplinary Design of Game-based Learning Platforms*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-04339-1

Ke, F., & Shute, V. J. (2015). *Serious Games Analytics*. (C. S. Loh, Y. Sheng, & D. Ifenthaler, Eds.), *Serious Games Analytics*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4

Keehn, S., & Claggett, S. (2019). Collecting Standardized Assessment Data in Games. *Journal of Applied Testing Technology*, *20*, 43–51.

Ketamo, H. (2013). Agents and Analytics - A Framework for Educational Data Mining with Games based Learning. In *Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (pp. 377–382). SciTePress - Science and and Technology Publications. https://doi.org/10.5220/0004331403770382

Ketamo, H. (2015). User-Generated Character Behaviors in Educational Games. In *Healthcare Informatics Research* (Vol. 21, pp. 57–68). https://doi.org/10.1007/978-981-287-408-5_5

Kickmeier-Rust, M. D. (2018). Predicting Learning Performance in Serious Games. In S. Göbel, A. Garcia-Agundez, T. Tregel, M. Ma, J. Baalsrud Hauge, M. Oliveira, ... P. Caserman (Eds.), *Serious Games* (Vol. 11243, pp. 133–144). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02762-9_14

Kitto, K., Whitmer, J., Silvers, A. E., & Webb, M. (2020). Creating Data for Learning Analytics Ecosystems. *SOLAR Position Paper*, 1–43.

Koedinger, K., McLaughlin, E., & Stamper, J. (2012). Automated Student Model Improvement. *Proceedings of the 5th International Conference on Educational Data Mining*, 17–24. https://doi.org/10.978.17421/02764

Kosmas, P., Ioannou, A., & Retalis, S. (2018). Moving Bodies to Moving Minds: A Study of the Use of Motion-Based Games in Special Education. *TechTrends*, *62*(6), 594–601. https://doi.org/10.1007/s11528-018-0294-5

Lazo, P. P. L., Anareta, C. L. Q., Duremdes, J. B. T., & Red, E. R. (2018). Classification of public elementary students' game play patterns in a digital game-based learning system with pedagogical agent. In *Proceedings of the 6th International Conference on Information and Education Technology - ICIET '18* (pp. 75–80). New York, New York, USA: ACM Press. https://doi.org/10.1145/3178158.3178160

Liu, M., Kang, J., Lee, J., Winzeler, E., & Liu, S. (2015). Examining Through Visualization What Tools Learners Access as They Play a Serious Game for Middle School Science. In *Serious Games Analytics* (pp. 181–208). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_8

Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning Analytics as an Assessment Tool in Serious Games: A Review of Literature. In *Serious Games and Edutainment Applications* (pp. 537–563). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-51645-5_24

Liu, M., Lee, J., Kang, J., & Liu, S. (2016). What We Can Learn from the Data: A Multiple-Case Study Examining Behavior Patterns by Students with Different Characteristics in Using a Serious Game. *Technology, Knowledge and Learning*, *21*(1), 33–57. https://doi.org/10.1007/s10758-015-9263-7

Loh, C. S., & Sheng, Y. (2014). Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior*, *39*, 322–330. https://doi.org/10.1016/j.chb.2014.07.022

Loh, C. S., & Sheng, Y. (2015a). Measuring Expert Performance for Serious Games Analytics: From Data to Insights. In *Serious Games Analytics* (pp. 101–134). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_5

Loh, C. S., & Sheng, Y. (2015b). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies*, *20*(1), 5–19. https://doi.org/10.1007/s10639-013-9263-y

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015a). *Serious Games Analytics*. (C. S. Loh, Y. Sheng, & D. Ifenthaler, Eds.). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015b). Serious Games Analytics: Theoretical Framework. In *Serious Games Analytics* (pp. 3–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_1

Long, P., & Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*(5), 30–32. Retrieved from http://search.proquest.com.proxy.library.vanderbilt.edu/docview/964183308/13AF5BC47C138E29FF2/5?accountid=14816%5Cnhttps://login.proxy.library.vanderbilt.edu/login?url=http://search.proquest.com/docview/964183308/13AF5BC47C138E29FF2/5?accountid=14816

Long, P., Siemens, G., Gráinne, C., & Gašević, D. (2011). LAK '11 : proceedings of the 1st International Conference on Learning Analytics and Knowledge, February 27

- March 1, 2011, Banff, Alberta, Canada. In *1st International Conference on Learning Analytics and Knowledge* (p. 195). Retrieved from https://dl.acm.org/citation.cfm?id=2090116

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.09.008

Manero, B., Torrente, J., Freire, M., & Fernández-Manjón, B. (2016). An instrument to build a gamer clustering framework according to gaming preferences and habits. *Computers in Human Behavior*, *62*, 353–363. https://doi.org/10.1016/j.chb.2016.03.085

Marchiori, E. J., Ferrer, G., Fernandez-Manjon, B., Povar-Marco, J., Suberviola, J. F., & Gimenez-Valverde, A. (2012). Video-game instruction in basic life support maneuvers. *Emergencias*, *24*(6), 433–437.

Martin, T., Aghababyan, A., Pfaffman, J., Olsen, J., Baker, S., Janisiewicz, P., … Smith, C. P. (2013). Nanogenetic learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (p. 165). New York, New York, USA: ACM Press. https://doi.org/10.1145/2460296.2460328

Martin, T., Petrick Smith, C., Forsgren, N., Aghababyan, A., Janisiewicz, P., & Baker, S. (2015). Learning Fractions by Splitting: Using Learning Analytics to Illuminate the Development of Mathematical Understanding. *Journal of the Learning Sciences*, *24*(4), 593–637. https://doi.org/10.1080/10508406.2015.1078244

Martinez-Garza, M. M., & Clark, D. B. (2017). Investigating Epistemic Stances in Game Play with Data Mining. *International Journal of Gaming and Computer-Mediated Simulations*, *9*(3), 1–40. https://doi.org/10.4018/ijgcms.2017070101

Mavridis, A., Katmada, A., & Tsiatsos, T. (2017). Impact of online flexible games on students' attitude towards mathematics. *Educational Technology Research and Development*, *65*(6), 1451–1470. https://doi.org/10.1007/s11423-017-9522-5

Mayer, I., van Dierendonck, D., van Ruijven, T., & Wenzler, I. (2014). Stealth Assessment of Teams in a Digital Game Environment. In *Lecture Notes in Computer Science* (Vol. 8605, pp. 224–235). https://doi.org/10.1007/978-3-319-12157-4_18

McCarthy, K. S., Johnson, A. M., Likens, A. D., Martin, Z., & McNamara, D. S. (2017). Metacognitive Prompt Overdose: Positive and Negative Effects of Prompts in iSTART. *Grantee Submission*, 404–405. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED577125&si

te=ehost-live

Michael, D. R., & Chen, S. L. (2005). Serious Games: Games That Educate, Train, and Inform. *Education*, *October 31*, 1–95. https://doi.org/10.1145/2465085.2465091

Mojang Studios. (2011). Minecraft. Retrieved from https://www.minecraft.net/

Mojang Studios. (2016). Minecraft Education Edition. Retrieved December 18, 2020, from https://education.minecraft.net/

Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Delgado Kloos, C. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*, pp. 1–1. https://doi.org/10.1109/TLT.2018.2856808

Muratet, M., Yessad, A., & Carron, T. (2016). Understanding Learners' Behaviors in Serious Games. In F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, & R. W. H. Lau (Eds.), *Advances in Web-Based Learning - ICWL 2015* (Vol. 9412, pp. 195–205). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47440-3_22

Nguyen, A., Gardner, L. A., & Sheridan, D. (2018). A framework for applying learning analytics in serious games for people with intellectual disabilities. *British Journal of Educational Technology*, *49*(4), 673–689. https://doi.org/10.1111/bjet.12625

Ninaus, M., Kiili, K., Siegler, R. S., & Moeller, K. (2017). Data-Driven Design Decisions to Improve Game-Based Learning of Fractions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10653 LNCS, pp. 3–13). https://doi.org/10.1007/978-3-319-71940-5_1

Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Evaluar el bullying y el cyberbullying validación española del EBIP-Q y del ECIP-Q. *Psicología Educativa*, *22*(1), 71–79. https://doi.org/10.1016/j.pse.2016.01.004

Owen, E., & Baker, R. (2019). Learning Analytics for Serious Games, (February). Retrieved from http://www.galanoe.eu/index.php/home/365-learning-analytics-for-serious-games

Owen, V. E., Anton, G., & Baker, R. (2016). Modeling User Exploration and Boundary Testing in Digital Learning Games. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16* (pp. 301–302). New York, New York, USA: ACM Press. https://doi.org/10.1145/2930238.2930271

Owen, V. E., & Baker, R. S. (2018). Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behavior Modeling in Serious

Games. *Technology, Knowledge and Learning*, (123456789). https://doi.org/10.1007/s10758-018-9393-9

Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. *International Journal of Artificial Intelligence in Education*, *24*(3), 251–283. https://doi.org/10.1007/s40593-014-0018-8

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*. https://doi.org/10.2753/MIS0742-1222240302

Pereira, H. A., De Souza, A. F., & De Menezes, C. S. (2016). A computational architecture for learning analytics in game-based learning. *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICALT 2016*, 191–193. https://doi.org/10.1109/ICALT.2016.3

Pérez-Colado, I., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2018). Game learning analytics is not informagic! In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1729–1737). IEEE. https://doi.org/10.1109/EDUCON.2018.8363443

Pérez-Colado, I. J., Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Simva: Simplifying the Scientific Validation of Serious Games. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (pp. 113–115). IEEE. https://doi.org/10.1109/ICALT.2019.00033

Pérez-Colado, I., Pérez-Colado, V., Martínez-Ortiz, I., Freire, M., & Fernández-Manjón, B. (2017). uAdventure: The eAdventure reboot - Combining the experience of commercial gaming tools and tailored educational tools. In *IEEE Global Engineering Education Conference (EDUCON)* (pp. 1754–1761). Retrieved from http://www.e-ucm.es/drafts/e-UCM_draft_304.pdf

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, *107*, 68–90. https://doi.org/10.1016/j.compedu.2017.01.004

Petrov, E. V., Mustafina, J., Alloghani, M., Galiullin, L., & Tan, S. Y. (2018). Learning Analytics and Serious Games: Analysis of Interrelation. In *2018 11th International Conference on Developments in eSystems Engineering (DeSE)* (Vol. 2018–Septe, pp. 153–156). IEEE. https://doi.org/10.1109/DeSE.2018.00037

Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., … Perlin, K. (2013). Metrics in Simulations and Games for Learning. In *Game*

Analytics (pp. 697–729). London: Springer London. https://doi.org/10.1007/978-1-4471-4769-5_31

Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology*, *8*(NOV), 1–16. https://doi.org/10.3389/fpsyg.2017.02029

Project Jupyter. (2020). Jupyter. Retrieved November 1, 2020, from https://jupyter.org/

Rahimi, S., Shute, V., Kuba, R., Dai, C.-P., Yang, X., Smith, G., & Alonso Fernández, C. (2021). The use and effects of incentive systems on learning and performance in educational games. *Computers & Education*, *165*, 104135. https://doi.org/10.1016/j.compedu.2021.104135

Roberts, J. D., Chung, G. K. W. K., & Parks, C. B. (2016). Supporting children's progress through the PBS KIDS learning analytics platform. *Journal of Children and Media*, *10*(2), 257–266. https://doi.org/10.1080/17482798.2016.1140489

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rowe, E., Asbell-clarke, J., & Baker, R. S. (2015). Serious Games Analytics to Measure Implicit Science Learning. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics* (pp. 343–360). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4

Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A. G., Barnes, T. M., … Edwards, T. (2017). Assessing implicit science learning in digital games. *Computers in Human Behavior*, *76*, 617–630. https://doi.org/10.1016/j.chb.2017.03.043

Ruiperez-Valiente, J. A. (2020). The Implementation Process of Learning Analytics. *Ried-Revista Iberoamericana De Educacion a Distancia*, *23*(2), 88–101. https://doi.org/10.5944/ried.23.1.26283

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education*, *23*(1–4), 94–114. https://doi.org/10.1007/s40593-013-0004-6

Seif El-Nasr, M., Drachen, A., & Canossa, A. (2013). *Game Analytics*. (M. Seif El-Nasr, A. Drachen, & A. Canossa, Eds.). London: Springer London. https://doi.org/10.1007/978-1-4471-4769-5

Seppala, T. J. (2016). CivilizationEDU takes the strategy franchise to school. Retrieved December 18, 2020, from https://www.engadget.com/2016-06-24-civilizationedu-takes-the-strategy-franchise-to-school.html

Serrano-Laguna, Á., Manero, B., Freire, M., & Fernández-Manjón, B. (2017). A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes. *Multimedia Tools and Applications*, *77*(2), 2849–2871. https://doi.org/10.1007/s11042-017-4467-6

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, *50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a little for big improvements: Application of learning analytics and videogames for student assessment. In *Procedia Computer Science* (Vol. 15, pp. 203–209). Elsevier.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2014). Application of Learning Analytics in educational videogames. *Entertainment Computing*, *5*(4), 313–322. https://doi.org/10.1016/j.entcom.2014.02.003

Serrano Laguna, Á. (2017). *Mejorando la evaluación de juegos serios mediante el uso de analíticas de aprendizaje*. Universidad Complutense de Madrid.

Sharples, M., & Domingue, J. (2016). Adaptive and Adaptable Learning. *Lecture Notes in Computer Science. Switzerland*, *9891*, 13–16. https://doi.org/10.1007/978-3-319-45153-4

Shoukry, L., Göbel, S., & Steinmetz, R. (2014). Learning Analytics and Serious Games: Trends and Considerations. In *Proceedings of the 2014 ACM International Workshop on Serious Games*. https://doi.org/10.1145/2656719.2656729

Shute, V. J., & Moore, G. R. (2017). Consistency and Validity in Game-Based Stealth Assessment. In *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. *The Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Shute, V., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of Research on Educational Communications and Technology: Fourth Edition* (pp. 311–321). https://doi.org/10.1007/978-1-4614-3185-5_3

Shute, V., & Ventura, M. (2013). Stealth Assessment. In *The SAGE Encyclopedia of Educational Technology* (p. 91). 2455 Teller Road, Thousand Oaks, California 91320: SAGE Publications, Inc. https://doi.org/10.4135/9781483346397.n278

Slimani, A., Elouaai, F., Elaachak, L., Yedri, O. B., & Bouhorma, M. (2018). Learning analytics through serious games: Data mining algorithms for performance measurement and improvement purposes. *International Journal of Emerging Technologies in Learning*, *13*(1), 46–64. https://doi.org/10.3991/ijet.v13i01.7518

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A Meta-Analysis of Data Collection in Serious Games Research. In *Serious Games Analytics* (pp. 31–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_2

Smith, S. P., Hickmott, D., Southgate, E., Bille, R., & Stephens, L. (2016). Exploring Play-Learners' Analytics in a Serious Game for Literacy Improvement. In T. Marsh, M. Ma, M. F. Oliveira, J. Baalsrud Hauge, & S. Göbel (Eds.), *Serious Games* (Vol. 9894, pp. 13–24). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45841-0_2

Snell, J., Atkins, M., Norris, W., Messina, C., Wilkinson, M., & Dolin, R. (2011). JSON Activity Streams 1.0. *Act. Streams Work.*, *22*(8), 2013.

Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). The Dynamical Analysis of Log Data Within Educational Games. In *Serious Games Analytics* (pp. 81–100). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_4

Stamper, J. C., Lomas, D., Ching, D., Ritter, S., Koedinger, K. R., & Steinhart, J. (2012). The Rise of the Super Experiment. *Proceedings of the 5th International Conference on Educational Data Mining*, 196–199. https://doi.org/10.1177/0003122412458508

Standford Medicine. (2013). SICKO. Retrieved April 24, 2018, from https://med.stanford.edu/news/all-news/2013/09/stanford-designed-game-teaches-surgical-decision-making.html

Steiner, C. M., Kickmeier-Rus, M. D., & Albert, D. (2015). Making sense of game-based user data: Learning analytics in applied games. *International Conference E-Learning*, 195–198. https://doi.org/10.1017/CBO9781107415324.004

Streicher, A., & Roller, W. (2017). Interoperable Adaptivity and Learning Analytics for Serious Games in Image Interpretation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10474 LNCS, pp. 598–601). https://doi.org/10.1007/978-3-319-66610-5_71

Streicher, A., & Smeddinck, J. D. (2016). Personalized and Adaptive Serious Games.

In R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, & K. Zweig (Eds.), *Springer* (Vol. 9970, pp. 332–377). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46152-6_14

Su, Y., Backlund, P., & Engström, H. (2020). Comprehensive review and classification of game analytics. In *Service Oriented Computing and Applications*. https://doi.org/10.1007/s11761-020-00303-z

Sucholutsky, I., & Schonlau, M. (2020). "Less Than One"-Shot Learning: Learning N Classes From M<N Samples. Retrieved from http://arxiv.org/abs/2009.08449

Tellioglu, U., Xie, G. G., Rohrer, J. P., & Prince, C. (2014). Whale of a crowd: Quantifying the effectiveness of crowd-sourced serious games. In *2014 Computer Games: AI, Animation, Mobile, Multimedia, Educational and Serious Games (CGAMES)* (pp. 1–7). IEEE. https://doi.org/10.1109/CGames.2014.6934151

Tlili, A., Essalmi, F., Jemni, M., & Kinshuk. (2016). An educational game for teaching computer architecture: Evaluation using learning analytics. *2015 5th International Conference on Information and Communication Technology and Accessibility, ICTA 2015*. https://doi.org/10.1109/ICTA.2015.7426881

United States Army. (2002). America's Army. Retrieved April 21, 2017, from https://www.americasarmy.com/

Vagg, T., Tan, Y. Y., Shortt, C., Hickey, C., Plant, B. J., & Tabirca, S. (2018). MHealth and Serious Game Analytics for Cystic Fibrosis Adults. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)* (Vol. 2018–June, pp. 100–105). IEEE. https://doi.org/10.1109/CBMS.2018.00025

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in Learning Analytics and Educational Data Mining. *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015 - Proceedings*, (January), 297–306.

Valve. (2011). Portal 2. Retrieved from https://store.steampowered.com/app/620/Portal_2/

Valve. (2012). Teach with Portals. Retrieved from http://www.teachwithportals.com/

Wallner, G., & Kriglstein, S. (2015). Comparative Visualization of Player Behavior for Serious Game Analytics. In *Serious Games Analytics* (pp. 159–179). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_7

Wang, T., Zhu, J.-Y., Torralba, A., & Efros, A. A. (2018). Dataset Distillation, 1–14. Retrieved from http://arxiv.org/abs/1811.10959

Wiemeyer, J., Kickmeier-Rust, M., & Steiner, C. M. (2016). Performance Assessment

in Serious Games. In *Serious Games* (Vol. 13, pp. 273–302). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40612-1_10

Xu, Y., Johnson, P. M., Lee, G. E., Moore, C. A., & Brewer, R. S. (2014). Makahiki : An Open Source Serious Game Framework for Sustainability Education and Conservation. *International Association for Development of the Information Society*, 8.

Xu, Z., & Woodruff, E. (2017). Person-centered approach to explore learner's emotionality in learning within a 3D narrative game. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17* (pp. 439–443). New York, New York, USA: ACM Press. https://doi.org/10.1145/3027385.3027432

Yang, X., Rahimi, S., Shute, V., Kuba, R., Smith, G., & Alonso Fernández, C. (2021). The relationship among prior knowledge, accessing learning supports, learning outcomes , and game performance in educational games (accepted). *Educational Technology Research & Development*.

Yannakakis, G. N., & Togelius, J. (2018). *Artificial Intelligence and Games*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-63519-4